EVALUATION AND ASSIGNMENT OF SIGNIFICANCE LEVELS TO PEPTIDE IDENTIFICATIONS FROM THE DATABASE SEARCH PROGRAMS USING RESAMPLING APPROACH

BY

MALIK NADEEM AKHTAR

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Animal Sciences
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Professor Sandra L. Rodriguez-Zas, Chair
Professor Jonathan V. Sweedler
Assistant Professor Maria B. Villamil
Professor Gustavo Caetano-Anolles

ABSTRACT

A novel application of Monte Carlo permutation testing that improves the calculation of the peptide match significance levels and detection rate in database search programs is demonstrated. Novel k-permuted decoy databases (where k denotes the type and number of permutations) were evaluated for accurate computation of match significance levels. K-permuted decoy databases were generated by: (a) complete permutations of peptide sequences (Whole), (b) permutation of terminal positions of peptide sequences (End), and (c) permuted peptides that fall within a certain mass tolerance of the tandem mass spectra (Mass-based). The 'Whole' and 'End' based permutation tests were performed using various indicators of peptide match quality in OMSSA, Crux, and X! Tandem on manually annotated neuropeptide tandem mass spectrometry spectra. Permutation *p-values* were calculated as the fraction of the permutations in the k-permuted databases with match indicator score as extreme as the original spectra match in the target database. The 'Whole' k-permuted decoy databases identified most (up to 100%) neuropeptides, while the 'End' k-permuted decoy databases provided better discrimination of the performance between the match indicators. The permutation test based *p-values* using the hyperscore (X! Tandem), *E-value* (OMSSA) and Sp score (Crux) match indicators outperformed the other match indicators in the database search programs. The simple indicator of match "the number of matched ions" provided performance comparable to the best match indicators in the OMSSA, X! Tandem, and Crux. Databases of least 10^5 k-permuted decoy peptides per spectra provided accurate *p-value*s. Overall, the 'Whole' and 'End' k-permuted decoy databases improved the consensus among the database search programs.

The ability of the k-permuted decoy databases to improve the classifications among correct and incorrect peptide matches was evaluated with 'Mass-based' k-permuted decoy database using best match indicator in the OMSSA (i.e., *E-value*). The evaluation was performed by searching 5806 tryptic tandem mass spectra (671 with annotated peptide entries) against the standard target and combined target-decoy databases. False discovery rate estimates based on the target-decoy approach and known identities of the annotated spectra were used to filter the peptide-spectrum matches. The k-permuted decoy database approach enabled the detection of up to 89% and 87% annotated peptides relative to the OMSSA's *E-value* with 82% and 84% identifications in the target database and target-decoy database, respectively. Improvements in performance was due to better performance of the k-permutation decoy database on small and large peptides with less than 13 matched fragment ions and large (insignificant) OMSSA *E-values*.

ACKNOWLEDGEMENTS

I owe my deepest gratitude to GOD Almighty who bestowed me with his blessings. I am very thankful to many people for their continuous support and inspiration throughout my graduate studies. I would like to thank my adviser Dr. Sandra L. Rodriguez-Zas, for her support, insightful comments and guidance, encouragement, patience, and her dedication during my entire Masters and Doctoral studies. I would like to thank Dr. Bruce R. Southey, whose passion, patience, support, advice, and humor made things easy for me more than I can even remember. I would like to thank my doctoral committee members, Dr. Jonathan V. Sweedler, Dr. Maria B. Villamil, and Dr. Gustavo Caetano-Anolles for their crucial advice, support, encouragement, and feedback during my preliminary and final examinations. I would like to thank Dr. Alfred Roca, for being always helpful, for being on my preliminary committee, and for acting as my research adviser in the Department of Animal Sciences during the annual graduate student evaluations. I would like to thank my lab members, Dianelys Gonzalez, Scott Nixon, Cynthia Zavala, Kelsey Caetano-Anolles, Kristin Delfino, and Robmay Garcia for their continuous support and friendship. I would like to thank my family and friends, for their love and support throughout the graduate studies. I would like to thank NIH (Grant Numbers: R21 DA027548, P30 DA018310 and R21 MH096030) for their support. I would like to thank COMSATS Institute of Information Technology (CIIT), for providing me an opportunity to become a part of international community of graduate students at University of Illinois Urbana-Champaign.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV

LIST OF TABLES

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER IV (CONT.)

CHAPTER I: LITERATURE REVIEW

1.1 NEUROPEPTIDES

Neuropeptides are a complex class of endogenous peptides containing both neurotransmitters and peptide hormones.[1] Neuropeptides perform multiple functions including communication between cells, and regulation of various biological processes such as growth, memory, learning, behavior, sleep, and circadian rhythms.[1, 2] Neuropeptides are present in the central nervous system and peripheral organs including pancreas, adrenal gland, and in the immune system.[3] Given the same primary amino acid sequence, some neuropeptides can act both as neurotransmitter and as a peptide hormone. Neuropeptides are functionally active molecules that are derived from larger inactive precursor proteins known as prohormones after complex proteolytical processing. A prohormone may contain one copy of the neuropeptide, multiple copies of the same neuropeptide, or multiple distinct neuropeptides.[1]

Most prohormones follow a common mechanism for the proteolytic processing.[1] Prohormones include an N-terminal signal peptide that guides the sequence through the ribosome and into the lumen of rough endoplasmic reticulum. Here, the signal peptide is cleaved by the signal peptidase enzymes followed by the transfer to the trans-Golgi apparatus. In the Golgi apparatus, the prohormones are packed into the secretory vesicles along with various processing enzymes.[4, 5] Formation of functionally active neuropeptides from prohormones in the secretory vesicles is a multi-step process. First, endoproteolytic cleavage by convertase enzymes generates intermediate neuropeptides. This cleavage occurs C-terminal from the dibasic or multiple basic residues (i.e., lysine or arginine), or less frequently from single basic residues, or rarely on from

non-basic residues.[6] Other factors that can influence the processing of prohormones into neuropeptides include the organism developmental stage and the environment such as pH.[7] Second, C-terminal basic residues are removed from the intermediate neuropeptides by the carboxypeptidases enzymes. Previous studies have shown that defects in the prohormone processing and failure to remove basic residues leads to obesity in humans and rodents.[8, 9] Third, the neuropeptides undergo further post-translational modifications (PTMs) including acetylation, phosphorylation, and amidation.[1, 10] **Figure 1.1** depicts the steps involved in neuropeptide processing. N- and C-terminal PTMs are the most common among neuropeptides and are important for optimal functional activity and low degradation of the neuropeptides.[6]

The resulting neuropeptides that are released into extracellular space are short in length, usually ranging between 3-40 amino acids.[1] Neuropeptides interact with G-protein coupled receptors located on the surface of the target cells. The receptors consist of seven membrane spanning alpha helices. The binding of the neuropeptide to the G-protein coupled receptors changes its conformation leading to activation of coupled G-protein, which then mediates intercellular signal transduction. So far approximately 100 different neuropeptide receptors have been reported in *C. elegans*.[11]

Several methods are available to identify neuropeptides from the biological samples including: Edman degradation, immunocytochemistry, enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA), and mass spectrometry (MS). Among these methods, MS has gained much popularity for the peptide and protein identification.

1.2 MASS SPECTROMETRY BASED PROTEOMICS AND PEPTIDOMICS

The disciplines of Proteomics and Peptidomics deal with the characterization of protein and peptide content within an organ, tissue or cell of the organisms, respectively.[12] MS is an analytical technique that has gained much popularity for the analysis of proteins and peptides present in the complex biological samples mainly due to improvement in separation techniques, availability of sequence databases, and soft ionization techniques (that transmit little residual energy onto the molecules to avoid too much degradation of molecules) such as electrospray ionization (ESI) and matrix assisted laser desorption ionization (MALDI).[13]

A mass spectrometer contains three regions: an ion source, a mass analyzer, and a detector region. The ion source converts proteins or peptides in a sample into ions for MS analysis. ESI and MALDI are the two common methods that vaporize the molecules out of solution and dry samples, respectively. ESI coupled with MS is most commonly used for complex protein mixtures, while for large number of relatively simple protein mixtures MALDI-MS is used. The mass analyzer region measures mass-to-charge (*m/z*) ratio of the ionized molecules. Various types of mass analyzers are available for the proteomic research that differs from each other in terms of sensitivity, mass resolution, mass accuracy, and ability to generate informative mass spectra. The basic types of mass analyzers include ion trap, time-of-flight, quadrupole, and Fourier transform ion cyclotron. The detector region determines the intensity value associated with each *m/z* value.[13]

Figure 1.2 shows an overview of the tandem mass spectrometry. A tandem mass spectrometer contains more than one mass analyzer regions that are separated by the collision chambers.[13] Upon injection of sample into the mass spectrometer, the ion source converts

molecules into ions that are analyzed by the first mass analyzer and an MS spectrum is generated that contains *m/z* values of the peptide ions and their relative abundance. The selected peptide ions undergo further fragmentation to generate a MS/MS or MS2 spectrum that contains information about the primary structure of the peptides.[13, 14] The downstream analysis of MS or MS/MS spectra provides information about the identity of the peptides or proteins. In MS-based analysis, characterization of the protein of interest is conducted either through bottom up approach or top down approach.[14, 15] Neuropeptides are endogenous peptides that are already present in the sample and do not require sample preparation by enzymatic digestion (bottom up approach) or MS-based fragmentation (top down approach) of proteins.[16] However, due to their typical short length the performance of the database search programs was tested on neuropeptides such that these peptides can be generated in the course of some protein experiments by protein digestion or fragmentation. Hence such approaches are described briefly.

TOP-DOWN APPROACH

In the top down approach, proteins in the complex mixture are separated and then intact proteins are subjected to ionization by ESI or MALDI. The ionized proteins are fragmented by MS to generate fragment ions. This provides molecular masses of both intact proteins and their fragment ions that can be used to identify protein of interest with more complete amino acid sequence coverage and information about PTMs.[15] For the top down approach, the fragmentation methods such as electron capture dissociation (ECD) and electron transfer dissociation (ETD) are more effective in fragmenting large peptides and proteins.[14, 15] Provided enough number of fragment ions are detected in MS for the protein of interest, the top down approach enables

identification of protein isoforms in much better fashion due to better sequence coverage.[14, 15] Furthermore, protein quantification using the top down approach is more reliable because abundance of proteins is measured directly rather than estimating it from the abundance of constituent peptides.[14] Another advantage of the top down approach is that the masses of intact proteins are dispersed over a wider mass range unlike the peptide mixture obtained from the enzymatic digestion of proteins, thus reducing the complexity associated with the requirement to separate peptides prior to MS/MS analysis.[17] Drawbacks of the top down approach include limitations associated with the separations methods, low sensitivity, and need for the large volumes of the sample relative to the bottom up approach.[14] The masses of intact proteins and their fragments ions are queried against the proteomic databases[18] or de novo approach can be used to identify the protein.[14]

BOTTOM-UP APPROACH

The bottom up approach is most commonly used to identify proteins present in complex biological samples in high throughput experiments. This approach starts with the protein purification step that is carried out either using gel based methods or gel free methods.[14] The separated proteins are enzymatically digested to generate complex set of peptides. Among several proteases, trypsin is most commonly used that digests proteins at carboxyl-terminus of arginine or lysine residues unless these are followed by the proline residue. The resulting peptides mixture is separated using single or multidimensional separation techniques. The separated peptides are ionized by MS using ESI or MALDI ionization sources to generate peptide ions.[15, 19] The mass analyzer region of the MS records the *m/z* values of the peptide ions (producing MS spectra).[13] In

bottom up studies mostly the peptide ions are further fragmented in tandem MS by Collision Induced Dissociation (CID) to generate product or fragment ions containing information about amino acid sequence and PTMs.[14] **Figure 1.3** depicts the general scheme of a typical bottom up experiment.

The bottom up approach has several advantages over the top down approach for the large scale protein identifications. The bottom up can deal with samples of high complexity and the peptides resulting from the enzymatic digestion are more easily separated than the intact proteins with the current front end separation techniques. Furthermore, bottom up needs lesser volume of the sample and is widely used for the quantification of peptides and proteins through chemical modifications of peptides with techniques such as ICAT or O^{18} labeling. However, quantification using the top down approach is more reliable.[14]

Several limitations are also associated with the bottom up approach. The digestion of proteins with enzyme such as trypsin results in peptides that fall within a relatively narrow mass range, which increases the difficulty to isolate these individual peptides for the downstream analysis.[17] Another challenge is the under sampling of peptides representing less abundant proteins and mostly proteins with high abundance are detected.[20] In the bottom up approach not all peptides from a single protein sequence are detected (usually 50-90% are detected) which leads to limited protein sequence coverage, which makes it less ideal choice to identify splice variants and PTMs.[21] Typically, only few peptides that provide sufficient information are used to identify the parent proteins in the bottom up analysis.[15]

Various computational methods are available to identify protein sequences using MS or MS/MS spectral data. For the MS scan, Peptide Mass Fingerprinting (PMF) can be used to identify the protein of interest by comparing masses of observed peptides with the masses of peptides generated from each protein sequence in the database. However, PMF is useful when the sample only contains pure proteins.[17] For the MS/MS data, the sequence database searching is the most efficient method to identify peptides. The magnitude of the correlation between the experimental and theoretical spectra in the database receives a statistical significance *p*- or *E-value*. Subsequently peptides are used to identify the precursor proteins by peptide-protein mapping and statistical confidence scores are assigned to peptide-protein mappings.[22] Several databases include information that can be used by the database search and spectral library search approaches.

1.3 DATABASES OF PROHORMONES AND NEUROPEPTIDES SEQUENCE AND SPECTRAL DATA

UNIPROT

UniProt (http://www.uniprot.org) is an integrated resource to store information pertaining to protein sequences and their functional annotation from various sources. The UniProt is a joint effort of research groups in the European Bioinformatics Institute (EBI), Protein Information Resource (PIR), and Swiss Institute of Bioinformatics (SIB).[23, 24] UniProt has four components: UniProt Knowledgebase (UniProtKB) is a central repository to store curated information about proteins along with cross-references to more than 140 databases providing additional or complementary information on the annotation.[24] UniProt Archive (UniParc) keeps track of changes

in the protein sequences present in the UniProt database. UniProt Reference Clusters (UniRef) clusters sequences in related species based on similarity to increase speed of searches. UniProt Metagenomic and Environmental sequences (UniMES) provide metagenomic and environmental data. The protein and peptide sequences from the UniProtKB were used in these studies.

The UniProtKB has two different components: UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The UniProtKB/Swiss-Prot contains manually curated protein sequences and annotations that are extracted from literature and computational analysis. For each protein the following information is provided: function, enzyme specificity, functional domains, PTMs, subcellular location, tissue specificity, spliced isoforms, structure, interactions, and associated diseases.[23] The current version of UniProtKB/Swiss-Prot (release 2013_10; October 16, 2013) contains 541,561 sequences obtained from 223,284 references. The UniProtKB/TrEMBL contains protein sequences that are computationally annotated and classified. The sequences that are translated from the coding sequences (CDS) present in the EMBL, GenBank, DDBJ nucleotide sequence databases, sequences associated with PDB structures, and data derived from the sequences directly submitted to UniProtKB and published literature.[23] Currently UniProtKB/TrEMBL (release 2013_10; October 16, 2013) contains 44,746,523 protein sequence records.

SWEPEP

SwePep (http://www.swepep.org) is a composite database of neuropeptide sequences and tandem spectral data designed to facilitate peptide identification in mass spectrometry

experiments.[25] The sequence database contains 4,180 annotated endogenous peptides and small proteins that are less than 10 KDa from 394 different species. These endogenous peptides were collected from three sources: in house experimentally verified peptides, peptides and proteins from UniProt, and peptides and proteins extracted from peer-reviewed publications. SwePep provides the calculated monoisotopic mass, average isotopic mass and isoelectric point (PI) for each peptide sequence in the database. The peptides in SwePep are classified into three groups: (1) biologically active peptides, the peptides with known biological functions; (2) potential biologically active peptides, the peptides with unknown biological function that belong to known peptide precursor proteins containing endogenous peptide specific processing sites; and (3) uncharacterized peptides, all peptides that do not belong to the above two groups. The SwePep sequence database is searchable by using the peptide's mass or name, organism name, or UniProt accession number.[25] The spectral library of the SwePep includes CID spectra obtained from the LTQ mass spectrometer coupled with liquid chromatography and ESI.[26] The 389 unique peptide identifications from 2,700 tandem spectra using X! Tandem were included in the spectral library regardless of the score threshold ($\log_e(-2)$). The spectral library is searchable using peptide sequence and peptide molecular mass with adjustable mass tolerance.

NATIONAL INTITUTE OF STANDARDS AND TECHNOLOGY (NIST)

NIST (http://peptide.nist.gov) hosts a tandem spectral library of tryptic peptides produced in LC-MS/MS experiments utilizing the ESI method. The library generally holds spectra from ion trap and quadrupole-TOF mass spectrometers. The tandem spectra in NIST are grouped into three categories: 1) the consensus spectra; 2) the best replicate spectra; and 3) the high confidence single

spectra identifications. Utilizing four database search programs the peptides were peptide sequences were assigned to every spectrum.[27]

NEUROPEDIA

NeuroPedia is a neuropeptide sequence and spectral library (http://proteomics.ucsd.edu/Software/NeuroPedia/index.html).[28] NeuroPedia was developed to improve the sensitivity and speed of the sequence database search and spectral library search programs in neuropeptide studies. The NeuroPedia sequence database contains 847 neuropeptides obtained from seven species, human (270), rat (195), mouse (188), bovine (154), rhesus macaque (20), chimpanzee (17), California sea hare (2), and leech (1). The 847 neuropeptides (from 332 precursor proteins) ranged in length from 2 to 1,129 amino acids in length. Neuropeptide description available in NeuroPedia includes peptide sequence, name, gene family, organism name, taxonomy, gene name, RefSeq gene identifier, protein name, RefSeq protein identifier, UniProt accession number, and start and end positions of the neuropeptide in the precursor protein. The 847 sequences are clustered into three groups regardless of the species based on sequence similarity in pairwise alignments: (1) 531 identical pairs, when the two aligned peptide sequences are exactly similar or redundant; (2) 5,020 overlapping pairs, with two aligned sequences identical up to half the length of the longest sequence in the alignment; and (3) 9,185 homolog pairs, with aligned sequences including one or two amino acid substitutions. The NeuroPedia spectral library contains 3,401 tandem spectra gathered from the NIST library and in house experimental datasets from five species, human (3,184), bovine (145), mouse (67), rat (4), and leech (1). The tandem spectral library is downloadable in the Mascot Generic Format (MGF) and is compatible with MSPLIT[29]

spectral library search program. The 3,401 tandem spectra are divided into ten MGF files depending upon the source organism (five species), instrument type (ion trap or quadrupole time of flight), and enzyme specificity (trypsin, v8, or none). Furthermore, the NeuroPedia allows visual inspection of each tandem spectrum.

PEPTIDEDB

PeptideDB (http://www.peptides.be) is a sequence database composed of biologically active endogenous peptides, precursor proteins and known protein motifs.[30] The current version of the database (version 1.0; April 25, 2008) contains 20,027 bioactive peptides derived from 19,438 precursor proteins obtained from 2,820 metazoan species. The peptides and proteins in the PeptideDB were collected from BLAST alignments, annotations in the UniProt database, and published literature. The 19,208 out of 19,438 precursor proteins in the PeptideDB were classified into 373 peptide families based on sequence similarities and information available in the literature, while the remaining 230 precursor proteins with no significant homology were classified in a "unique peptide group". The 48% (178) of protein families have known motifs in the Prosite (http://prosite.expasy.org), Pfam (http://pfam.sanger.ac.uk), SMART (http://smart.embl-heidelberg.de), and peptidemotifdat[31] databases. The peptide and precursor protein length distribution indicated that 97% peptides and 98% precursor proteins are less than 200 and 500 amino acids in length, respectively. The PeptideDB database is searchable using the PeptideDB accession number, peptide name, peptide length, monoisotopic mass, amino acid sequence, organism common name, peptide family, or UniProt accession number.

PEPSHOP

PepShop (http://stagbeetle.animal.uiuc.edu/pepshop.html) is a comprehensive web resource that enables the identification and discovery of neuropeptides.[32] PepShop integrates public databases encompassing sequence, annotation, and tandem mass spectra (MS/MS) information with bioinformatics and proteomics tools to input, search, align, predict, and identify prohormone and peptides. PepShop integrates experimentally confirmed prohormone and peptide information from the SwePep, UniProt, and NeuroPred repositories. The PepShop data warehouse can be searched by species (seven species), prohormone identifier (668 unique sequences), exact amino acid sequence, and peptide monoisotopic mass with adjustable mass tolerance level. The neuropeptides in the PepShop database are linked to the spectral library of SwePep. PepShop enables the search of user-provided MS/MS profiles against the in-house neuropeptide repository using three open source database search programs, Crux, X! Tandem, and OMSSA. In PepShop, identified peptides are automatically linked to prohormone and peptide information.

NEUROPEPTIDES.NL

Neuropeptides.nl (http://www.neuropeptides.nl) database contains information about the known neuropeptides, neuropeptide genes, precursor proteins, and their expression in the mouse brain.[33] The neuropeptide genes have been grouped into families based on structural or functional similarities among them. The neuropeptide genes are linked to their corresponding locus on the human genome through UCSC (University of California Santa Cruz; http://genome.ucsc.edu) human genome browser. The UCSC browser provides further information about the gene location,

transcripts, and base wise conservation in other species. The precursor proteins are linked to their isoforms and homologous proteins in related species using pre-computed BLAST results. Comparisons of precursor proteins across species indicate that precursor proteins are less conserved relative to the neuropeptides and their processing sites. Furthermore, the neuropeptide genes are linked to the mouse expression data for the annotated genes in the online Allen Brain Atlas or GenePaint.org resources.

EROP-MOSCOW

EROP-Moscow (http://erop.inbi.ras.ru) database provides comprehensive information about 10,575 naturally occurring bioactive oligopeptides.[34] These peptides ranged from 2 to 50 amino acids in length. Of 10,575 bioactive peptides in the current version of the database, 2,362 peptides are neuropeptides. The database provides information about each neuropeptide including peptide length, sequence, precursor protein, PTMs, biological functions, molecular mass, isoelectric point, and literature sources. The majority of the information about neuropeptides and other functional classes of bioactive peptides (such as toxins, antimicrobial) was extracted from the scientific literature. The peptides are also linked to the external generalized databases, Swiss-Prot, protein identification resource (PIR), and PubMed. Based on sequence similarity peptides are also grouped into homologous families.

NCBI REFSEQ

The National Center for Biotechnology Information (NCBI) Reference Sequence (http://www.ncbi.nlm.nih.gov/RefSeq/) is a collection of genomic, transcripts and protein sequences. The database contains more than 13 x 10^6 protein entries from more than 16,000 species.[35] RefSeq contains well annotated sequences for the neuropeptide genes and precursor proteins. The key features of the RefSeq are less redundancy in records and improved cross-referencing between nucleic acid and protein information. The RefSeq records are generated either using annotation pipelines or through manual annotation. The accession numbers of the protein records derived from the annotation pipelines and manual annotation are denoted with prefixes "XP_" and "NP_", respectively.

1.4 PEPTIDE IDENTIFICATION BY TANDEM MASS SPECTROMETRY

Several computational approaches and software tools are available to identify peptide sequences from tandem spectra. These approaches are grouped into four categories depending upon how peptide sequence is assigned to the tandem mass spectra: (1) *de novo* peptide identification, (2) sequence database searching, (3) spectral library searching, and (4) hybrid approach.

The *de novo* approach extracts peptide sequences from the experimental spectra without any prior knowledge about the peptide sequences.[36] This approach is based on the rationale that the two fragment ion peaks in the tandem spectra differ by a single amino acid and sequence of the peptide can be obtained by calculating the mass differences between the adjacent peaks.[37] Novel

peptides can be identified by this approach but at the same time the error rate is high due to incomplete fragmentation patterns in tandem mass spectra.[36, 38]

The database search approach identifies spectra by comparing experimental spectra against theoretical spectra generated from peptides in the sequence database. This approach is useful when the peptide sequence is known and present in the sequence database or when the experimental spectra have low quality and incomplete fragmentation. The sequence database approach can match tandem mass spectra containing sufficient information (i.e., signal peaks) to peptide sequences in the database even if the spectra are of poor quality (too many non-signal peaks or low intensity of signal peaks) and contains incomplete fragmentation.[37] However, confidence in peptide identification is decreased if the spectrum quality is too low[39] or when many fragment ions are missing.[16] The database search approach cannot identify those peptides that are not present in searched database.[22]

The spectral library approach identifies peptides by searching the experimental spectrum against already annotated spectra present in the spectral library.[22, 40, 41] The spectral library search approach is based on the rationale that MS-based peptidomics experiments include many peptide spectra already annotated in prior studies.[42] Like the sequence database search approach, this approach cannot identify novel peptides.

The hybrid approach is a combination of *de novo* and sequence database search approach. In the first step, short sequence tags (i.e., short sequences of 3-5 amino acids in length) are extracted from the tandem mass spectra using *de novo* approach and then these sequence tags are searched against the sequence databases using database search methods. This approach is designed

to overcome the limitations of *de novo* approach (failure to correctly match spectra with incomplete information and poor quality) and the database search approach (identification of novel or mutated peptides).[22, 41] The database search programs were used in the current studies and will be discussed in detail.

1.5 OVERVIEW OF THE DATABASE SEARCH APPROACH

The database search approach is the most common approach to detect peptides in the bottom up proteomics studies primarily due to the ability to handle spectra with incomplete fragmentations and of low quality (low intensity of the signal peaks or presence of many non-signal peaks). **Figure 1.4** shows an overview of the database search approach. The database search programs correlate experimental spectra with the *in silico* theoretical spectra generated from the peptide sequences in the database.[40] One or more scores or indicators are reported with each score indicating the strength of the peptide-spectrum match. Furthermore, in addition to existing proteomic databases, this approach can use information from translated genomic databases.[22, 41]

Many database search programs have been developed and routinely used for the peptide identification including OMSSA,[43] Crux,[44] X! Tandem,[45] Mascot,[46] SEQUEST,[47] and Tide.[48] These programs differ in the heuristic search algorithms and the way the experimental-theoretical spectra matching score is computed.

The database search programs match individual spectrum against a subset of all the peptides present in the sequence database that fall within the mass range (tolerance) of the precursor peptide. The scores are converted to either *p-values* or *E-values* that reflect the statistical significance of the match based on a theoretical test distribution or an empirical test distribution

based on other peptide spectrum matches.[41] The *p-value* is the probability of obtaining the match between experiment and theoretical spectra due to chance. The *E-value* is closely related to *p-value* but denotes the expected number of random database matches that received score as high as the current match.

SEARCH PARAMETERS FOR THE DATABASE SEARCH PROGRAMS

The parameter specification of the database search program affects peptide identification. The parameters influence the selection of the candidate peptides that have similar mass as the experimental spectrum, peptide identification accuracy, and speed of search.[22, 49] There is no best set of parameter values and the optimal search parameter values depends on multiple factors including tandem MS datasets, search methods and tools, and analysis strategies.[49] The most widely used search parameters are: monoisotopic or average isotopic mass, precursor and fragment ion tolerance, enzyme specificity, PTMs, and type of fragment ions.

Monoisotopic or Average mass

All database search programs allow the specification of the method to calculate the peptide masses from the *m/z* values of the peptide ions. The calculated masses from the *m/z* values can be closer to the monoisotopic mass (with ^{12}C atoms only) or average mass (including ^{13}C atoms). The monoisotopic mass is the mass of the most common isotopic form of the amino acids, while the average mass represents the weighted average of all the isotopic forms of the amino acids. The monoisotopic and average isotopic masses are usually used for the high-resolution and low-resolution mass spectrometers, respectively.[22]

Precursor and fragment ion tolerance

After calculation of the peptide mass from the *m/z* value of the precursor ion from the experimental spectrum, the database search program selects the database peptide sequences (candidate peptides) that fall within a certain mass range (precursor ions tolerance) of the experimental spectrum. The choice of the precursor mass tolerance value depends on the accuracy of the mass spectrometers that range from 0.05 Da for the high mass accuracy instruments such as Fourier transform to 3 Da for the low mass accuracy instruments such as ion traps.[22] The higher value of the precursor ion tolerance can affect the speed of searches and accuracy of peptide identification due to large number of available candidate peptides.[41] However, studies have shown that selection of few candidate peptides can also hamper the performance of the database search programs.[50] This is because many database search programs use the score distribution of the candidate peptides to assign significance values to the correct peptide match. The lack of sufficient candidate peptides can lead to potential incorrect matches. In addition to the precursor ion tolerance, the fragment ion tolerance can also be provided for the database search programs.

Enzyme specificity

The choice of the digestion enzyme to process the protein sequences into peptides depends on the experimental settings. Accurate specification of the digestion rules can reduce the search space to only those candidate peptides that satisfy the digestion rules of interest. Most database search programs are designed for tryptic peptides; however these programs can also be used for the neuropeptide searches by specifying custom cleavage rules. For neuropeptide searches, the protein library can be processed with a nonspecific enzyme that cleaves on every peptide bond while allowing for the large number of missed cleavages.[26] An alternative strategy is to instruct the

database search program to use the peptide sequence database without further processing.[16] NeuroPred and similar tools can be used to create such peptide databases.

Post-translational modifications (PTMs)

PTMs are the covalent modifications in the proteins that occur either due to proteolytic cleavage or addition of modifying groups.[51] So far, approximately 200 different types of PTMs have been reported.[52] Each modification makes the mass of the precursor and fragment ions different from the masses of peptides in the sequence databases. The database search programs select candidate peptides from the sequence database on the basis of observed mass and failure to incorporate these PTMs would lead selection of incorrect candidate peptides.[53] Most database search programs allow the specification of three different types of PTMs: (1) the modification of specific residue when present at peptide terminus such as pyro-glutamination of glutamine and glutamic acid residues; (2) modifications of any residue present at peptide terminus such as N-terminal acetylation and C-terminal amidation; and (3) modification of particular residues regardless of their position in the sequence such as phosphorylation of serine, threonine, and tyrosine.[46] The PTMs can be applied either in fixed fashion (all occurrences of the residue are modified e.g., addition of 57 Da on every occurrence of cysteine due to cysteine alkylation) or in variable fashion (residue is only conditionally modified). The variable modification increases the search space exponentially with increase in the number of PTM specified, which can lead to reduction in search speed and peptide identifications.[20] Common PTMs for the neuropeptides are glycosylation, amidation, acetylation, phosphorylation, and sulfation. These PTMs occur in secretory granules and are species- or tissue-specific.[4]

Types of fragment ions

The value of this parameter depends on the type of fragmentation method used in the mass spectrometry.[22] The ions are named based upon the type of bonds that are broken between the two adjacent amino acids during the fragmentation process. The most common fragmentation is the CID which produces *b*- and *y*-ions due to the breakage of amide bonds. The breakage of bond between the alpha carbon and carbonyl carbon yields *a*- and *x*-series ions. Furthermore, the methods such as ETD mainly results in *c*- and *z*-ions due to the fragmentation of bonds between the amide nitrogen and alpha carbon. The fragment ions are classified as N-terminal or C-terminal if the charge is retained on the N-terminus or C-terminus of the peptide, respectively. The N-terminal ions include *a*-, *b*-, and *c*-ions, while *x*-, *y*-, and *z*-ions are classified as C-terminal ions.[54, 55] The database search program predicts the fragment ions for the selected candidate peptides according to this search parameter and then compares them with the fragment ions present in the experimental spectrum.[49]

1.6 REVIEW OF SELECTED DATABASE SEARCH PROGRAMS

Many database search programs are available including OMSSA, Crux, Mascot, Sequest, Tide, Myrimatch, and X! Tandem. A brief description of the selected database search programs, their scoring schemes and conversion of scores to either *E*- or *p-value* is given below. The X! Tandem and OMSSA use a parametric approach (fitting parametric distributions without using decoy peptides) to obtain significance values, while Crux uses a semi-supervised parametric approach (fitting parametric distributions from the scores of decoy peptides) to compute the *p-value.*

X! Tandem

X! Tandem (http://www.thegpm.org/tandem) is an open source program written in the C++ programming language and can be executed in multiple platforms (Windows, Linux, OS X).[45] X! Tandem assigns peptide sequences to provided tandem spectra in the multistep process. First, X! Tandem preprocesses the input tandem spectra to remove noise and artifacts (i.e., peaks resulting from the ions other than the selected ions in MS) using information provided in the search parameter file. The X! Tandem selects the 50 (user adjustable) most intense fragment ion peaks and the intensity values of the selected peaks are normalized using a user-adjustable dynamic range value (a parameter showing the difference between the most intense and least intense fragment peak in the spectra; the default value is 100). In the normalization step, the intensity of the most intense peak is set to one-hundred, while the intensities of the remaining peaks are linearly scaled with respect to most intense peak. Furthermore, peaks with scaled intensity below one are removed from the normalized spectrum. Second, X! Tandem processes the database protein sequences into peptides using specified enzymatic cleavage rules and the resulting peptide sequences are further subjected to chemical and PTMs. Third, the normalized observed spectrum is correlated to the theoretical spectra generated from the peptide sequences from the target search database. This step assigns scores to each peptide-spectrum match indicating the strength of the match.[45] Fourth, X! Tandem creates an XML output file containing details of the match such as precursor ion mass, charge state, hyperscore, *E-value*, peptide sequence, protein sequence, search parameters and others (http://www.thegpm.org/ docs/X_series_output_form.pdf).

X! Tandem first computes a convolution score (preliminary) for each peptide-spectrum match. The convolution score is the dot product of the intensities of the matched fragment ions

between experimental and theoretical spectra. The dot product is used because only the matched ions are considered. The convolution score is converted into a hyperscore by multiplying the score by the factorial number of matching *b*- and *y*-ions (the usage of *b*- and *y*-ions corresponds to the CID spectra). The default use of the factorial of the number of matched *b*- and *y*-ions can be modified to include other ions such as *a*-, *c*-, *x*- and *z*-ions in the scoring. The use of factorial is based on the hypergeometric distribution. The hyperscore is calculated as:

$$S_{hyperscore} = (n_b! \, n_b!) \sum_i x_i \, y_i$$

The database search produces a hyperscore distribution of all the peptide-spectrum matches, which is assumed to follow a hypergeometric distribution. The hypergeometric distribution is a parametric discrete probability distribution that allows extrapolation. The hyperscores are log-transformed and the hyperscores higher than the intersection between the log-transformed hyperscores (on the *x*-axis) and log transformation of the frequency of the hyperscores (i.e., *E-value* on the *y*-axis) are assumed to be significant.[39]

OMSSA

Open Mass Spectrometry Search Algorithm (OMSSA; http://pubchem.ncbi.nlm.nih.gov/omssa) is an open source program written in the C++ programming language that can be compiled across multiple platforms including Windows, Linux, Solaris, and OS X.[43] OMSSA uses a multi-step strategy to identify peptides from the spectra.

In the first step, OMSSA determines the precursor charge state of the spectrum by counting the number of peaks that fall below the *m/z* value of the precursor ion. A spectrum with more than 95% peaks below precursor *m/z* is considered in +1 precursor charge state, while the spectrum is

searched with +2 and +3 precursor charge states if less than 95% peaks fall below precursor *m/z* values. The accurate determination of the charge state is important in OMSSA because candidate peptides (peptides within the precursor mass tolerance) for each spectrum are selected using the neutral mass (i.e., sum of the masses of amino acid residues in a peptide and mass of the hydroxyl group) of the precursors. The second step involves preprocessing of experimental spectra to remove noise peaks including peaks with intensity below 2.5% of the highest peak in spectrum, precursor ion peaks, peaks that are within 2 Da of *m/z* distance from the examined peaks, and peaks that can be explained by neutral mass losses (loss 17 Da for ammonia, and 18 Da for water). Furthermore, peaks are examined in the order of intensity, for the precursor charge states +1 and +2 only the most intense peak within ±27 Da of the peak being examined is selected, while for the +3 charge state the two most intense peaks are selected within ±14 Da of the peak being examined. Third, candidate peptides from the sequence database that fall within precursor mass tolerance of the spectra are selected. The candidate peptides masses are calculated considering the specified PTMs. Fourth, to improve the speed of searches, the *m/z* values from the experimental spectra are converted to integer values using 100 as the scaling factor (user-adjustable), the sequence library is mapped to memory, and the observed spectra are sorted, and indexed by the precursor mass. Fifth, +1 charge fragment ions are calculated for precursor charge states +1 and +2, while both +1 and +2 fragment ions are calculated for precursor charge state +3 when the peak is above *m*/2 and below *m*/2, respectively, where *m* is the precursor mass. Sixth, the fragment ion peaks in the experimental and theoretical spectra are compared and a score is calculated. Only the theoretical spectra that have at least one fragment ion match with any of the top three (user-adjustable) most intense peaks in the experimental spectrum are scored to improve the sensitivity of the algorithm.[43]

The scoring of the experimental-theoretical spectra matches is based on the assumption that the distribution of the number of matched ions follows a Poisson distribution. Lambda (the Poisson mean parameter) is calculated by considering the fragment ion tolerance, number of peaks in the experimental and theoretical spectra, and the mass of the precursor. The lambda is calculated by counting the number of spectrum peaks that fall within two matched fragment ions from any one fragment ion series (e.g., *b*- or *y*-ions). The count is adjusted by dividing with the mass of the precursor. Lambda is the sum of the adjusted counts. The calculation of the lambda parameter is different for the spectra with +1 charge fragment ions than spectra with both +1 and +2 fragment ions. During the preprocessing step, the OMSSA noise filter removes some but not all noise peaks (peaks not representing fragment ion peaks are known as noise peaks) from the experimental spectra leading to inclusion of noise peaks in the calculation of the Poisson mean. The probability of the match with a given number of fragment ion matches (*x*) and lambda can be calculated as follow:

$$P(x,\mu) = \frac{\mu^x}{x!} e^{-\mu}$$

The OMSSA report results according to *E-value* which is the expected number of random database matches with probability equal or more significant than the one observed due to chance. OMSSA calculates this *E-value* by multiplying the number of candidate peptides (i.e., database peptides within precursor tolerance of observed spectra) with Poisson probability of the match.

$$E(y,\mu) = \mathrm{N}(1 - (\sum_{x=0}^{y-1} P'(x,\mu_Z))^N)$$

Crux

Crux is an open source reimplementation of Sequest, the first commercial database search program.[44] Like in the previous database search programs, the first step in Crux is the identification of all database candidate peptides that are within the precursor mass tolerance range of the experimental spectra. The candidate peptides are selected either by querying spectrum masses against the entire sequence database or against an indexed database of predicted peptides. This indexed database is a preprocessed binary sequence database obtained from the *in silico* digestion of precursor sequences in the target database and followed by sorting of the resultant *in silico* generated peptides by their masses. The index database allows efficient retrieval of candidate peptides upon query allowing Crux to perform faster searches than the original Sequest program. The candidate peptides are matched with the experimental spectrum and indicators of the strength of the experimental and theoretical spectra matches are reported. These indicators are: cross-correlation score (XCorr), delta Cn (ΔCn), Sequest preliminary score (Sp), and *p-value*.

First, Crux processes the spectra by taking the square root of each intensity peak value, normalizes the peak intensities to sum to one hundred, and round each *m/z* to the nearest integer value. Second, Crux uses the 200 most intense peaks to compute a Sp score. The higher the value of Sp score denotes higher similarity between theoretical and experimental spectra. The Sequest version used Sp score to filter top 500 database candidate peptides that are subsequently scored and reranked using XCorr to increase the speed of searches. The default version of Crux "search-for-matches" does not calculate Sp score. In this study, the Crux parameter file was modified to retrieve this score.

The experimental spectrum is preprocessed prior to calculating XCorr scores, the primary score of Crux that indicates the similarity between experimental and theoretical spectra. First, spectra are processed by taking the square root of each intensity peak value and rounding each *m/z* to the nearest integer value. The processed spectrum is divided into ten bins and the peaks intensity in each bin is set to a maximum of 50. A theoretical spectrum is synthesized for each candidate peptide containing *b*- and *y*-ions with peak intensity of 50, ±1 *m/z* peaks with intensities of 25, and with peak intensity 10 for *b*- and *y*-ion peaks with neutral mass loss of ammonia and *b*-ion peaks with neutral loss of water. The two spectra are correlated and higher XCorr denotes higher similarity between experimental and theoretical spectra. Crux computes a relative score (ΔCn) from the XCorr scores denoting the relative ranking of each peptide match in terms of other peptide matches for any particular spectrum. The deltaCn reflects the difference in the XCorr score of the top peptide-spectrum match relative to other matches for that spectrum. The deltaCn score indicates the strength of the top match relative to the second best match.

Crux calculates a *p-value* from a Weibull distribution obtained by using XCorr scores from all peptide-spectrum matches.[56] The *p-value* is the probability that the match between the experimental and target peptide spectrum is due to chance. Crux reports spectrum specific Bonferroni-adjusted *p-values*, adjusted by the number of candidate peptides. Crux uses 40 Weibull points (the minimum numbers of XCorr scores required to estimate the *p-value*) to estimate *p-values*. However, prior studies have shown that increase in the number of Weibull points increase significance levels of the estimated *p-values*.[16] Crux generates in silico peptides (decoy peptides described in the next section) when the number of candidate peptides are less than the number of Weibull points required to estimate *p-values* using Weibull distribution. The decoy peptides are

generated by keeping the terminal amino acids of the candidate peptides fixed while shuffling the internal amino acids of the peptide. The Crux sampling with replacement procedure is repeated until a minimum number of Weibull points are obtained. The source code of Crux was modified to obtain raw *p-values* for each spectrum using different values of Weibull points.

1.7 FACTORS AFFECTING PEPTIDE IDENTIFICATION

Accurate peptide identification from the tandem spectra remains challenging despite the many parametric, semi-parametric, and non-parametric methods available to calculate the significance levels of a match between the experimental and theoretical spectra. The significance levels provide an objective criterion to assess the likelihood that the scores of target peptides could be observed by chance. A large number of observed spectra are either missed due to significant levels that do not surpass the minimum user-defined threshold (false negative) or the match significance surpasses the minimum threshold yet the match is incorrect (false positive). Several factors influence the significance levels of the peptide-spectrum matches including: search space, peptide length, missing ions and low spectrum quality, and incomplete databases.

IMPACT OF THE SEARCH SPACE DENSITY ON PEPTIDE IDENTIFICATION

The precursor mass tolerance, choice of a digestion enzyme, and PTM searches influence the effective database size (the number of database peptides that have mass within the precursor tolerance of the experimental spectrum).[41, 49] A spectrum with fewer database candidate peptides is more likely to produce a correct peptide match relative to a spectrum with more candidate peptides. The presence of large number of candidate peptides reduces the sensitivity of the database search

programs as more incorrect peptides have a chance to receive a score higher than the correct peptide matches leading to an increase in the false positive results. The higher number of incorrect peptides with score as extreme as the correct target peptide leads to lower significance values for the corresponding spectrum. On the other hand, many database search programs use all candidate peptides scores for a spectrum to fit a distribution and calculate significance values for the match.[49] For example, X! Tandem estimates *E-value* from the distribution of hyperscores from all peptide matches for a spectrum.[57] Low number of candidate peptides increases the *E-value* and the match becomes less significant.[16] A wider precursor tolerance can be used to generate enough number of candidate peptides to estimate significance values in the absence of sufficient candidate peptides.[49] An alternative approach is to generate decoy peptides when sufficient number of candidate peptides is not available to estimate significance.[56]

IMPACT OF PEPTIDE LENGTH ON PEPTIDE IDENTIFICATION

The ability of the database search programs to accurately identify peptides mainly depends on the availability of a sufficient number of matching fragment ions.[16] Short peptides have a higher chance to be missed by the database search programs due to less significance values.[16, 20, 58] The short peptide tends to receive a score that is not different from the other matches of the spectrum due to less number of possible fragment ions. This problem is further complicated by other factors such as missing ions and presence of fragment ions due fragmentation of more than one peptides in a single spectrum.[16]

IMPACT OF MISSING IONS AND LOW SPECTRUM QUALITY ON PEPTIDE IDENTIFICATION

The incomplete fragmentation and noise (non-signal spectra peaks) in the spectra reduces the number of correct peptide identifications due to lower significant *p*- or *E-value* levels assigned by the database search programs. The increase (becoming less significant) in *p*- or *E-values* with both factors is due to the lower score of the correct matches that is not significantly different from the other matches of the spectrum. Most database search programs use intensity to select signal peaks to be used in their scoring functions. The low intensity of signal peaks relative to the noise peaks reduces the contribution of signal peaks in the scoring functions which can lead to the lower scores for the correct peptide matches.[26] In the case of same PTM occurring on more than one residue on a single peptide (e.g., phosphorylation of serine and threonine), the confidence in localization of the PTM is reduced in the absence of fragment ions representing the exact residue modified.[49]

FACTORS IMPACTING SEQUENCE DATABASES

The database search programs can fail to identify a peptide match if the corresponding peptide sequence is absent from the target database. This could be either due to the: (a) presence of a closely related variant of the peptide or protein rather than exact sequence in the database; (b) sequencing errors; (c) mutation in the sequence; (d) polymorphism; or (e) presence of homologous sequence in the database from closely related species.[59]

In the context of the standard database search approach (searching database with a narrow precursor mass tolerance) a peptide sequence is considered missing when either a peptide sequence

is totally absent from the sequence database or a closely related variant of the sequence is present in the database. This is because it can change the peptide mass and the resulting MS/MS fragmentation patterns of the *b*- and *y*-ions making the observed-theoretical spectra unmatchable. The error tolerant searches are assumed to work better than the standard database searches in such cases.[59]

Another reason could be the complex dissociation chemistry of peptides in MS that can permute or rearrange the sequence of peptides in sample. Studies have shown that the larger *b*-ions have higher tendency to form cyclic structures in which sequence ends are fused together followed by reopening of the ions at different residues instead of the original fused positions. Most database search programs do not take into account the possibility of peptide ion rearrangements while counting the number of shared peaks between the observed and theoretical spectra.[60] The exclusion of such permuted ions from scoring can contribute towards lower scores for the peptide matches.

MULTIPLE HYPOTHESIS TESTING

Typical MS-based peptidomics or proteomics experiments involve the analysis of thousands of experimental tandem spectra leading to a multiple hypothesis testing scenario.[61] Two different types of measures have been proposed to control multiple hypothesis testing problems: family wise error rate (FWER) and false discovery rate (FDR). The first measure is FWER the probability of rejecting at least one true null hypothesis among all *m* independent hypotheses. Given *m* independent hypothesis and the probability of error for each test (α), the FWER is calculated as:

$$FWER = 1 - (1 - \alpha)^m$$

The quantity *(1- α)* represents the probability of no error and *(1- α)*^{m} represents the probability of no error in the *m* independent tests.[62] For example for *m* = 100 and α = 0.05 then FWER is 0.99 (or 99% chance of observing at least one falsely rejected null hypothesis). Many methods have been proposed to control such high FWER and these methods are divided into categories: (a) single step approach, in which all *p-values* are adjusted equally; and (b) sequential step approach, in which each *p-value* is adjusted separately.

The Bonferroni adjustment is a single step approach that is used in the current study. This adjustment can be applied in two ways: (1) by multiplying the probability of type I error (alpha level) by the number of tests (adjusted alpha level) and accepting or rejecting the null hypothesis by comparing significance values against adjusted alpha level; and (2) by adjusting the raw significance values by multiplying them by the number of hypothesis tests and then comparing the adjusted significance values against the alpha level to accept or reject the null hypothesis. The Bonferroni adjustment is a highly conservative approach.

An alternative approach to control FWER is the Holm's sequential step wise adjustment method.[63] In this method the unadjusted *p-values* are arranged in an ascending order (from the smallest to the largest) and each unadjusted *p-value* is adjusted by multiplying with *m-j+1*, where *m* refers to the total number of tested hypotheses and *j* is the rank of the unadjusted *p-value* in the ordered list. The Holms method is less conservative than the Bonferroni method and the hypothesis rejected by the Bonferroni method would also be rejected in the Holm's step down procedure.

Benjamini and Hochberg proposed a false discovery rate (FDR) method as a second measure to handle multiple testing problems.[64] This method allows a certain percentage of false positive hypotheses among all rejected hypothesis. The FDR is defined as the expected fraction of false positive identifications or hypothesis among all rejected hypotheses. The FDR is calculated at certain threshold (α) by dividing the number of false positives (FP) with the total number of rejected hypothesis i.e., true positives (TP) and false positives (FP). Thus, peptide-spectrum matches need to reach statistical significance values that surpass the stringent threshold that controls for multiple hypothesis testing.

1.8 FDR VIA TARGET-DECOY APPROACH

In the MS/MS-based peptidomics studies, the FDR can also be calculated using the target-decoy database search strategy[41, 65-68] or mixture model approach. The target-decoy approach (TDA) is the simplest and most popular approach to estimate error rate. This approach is easily applicable to several experimental setups and demonstrates the ability of the scoring functions to distinguish between correct and incorrect peptide-spectra identifications. The TDA is based on the assumption that the score distribution of incorrect matches from the target database is identical to the score distribution of the decoy matches.

The accuracy of the TDA based FDR estimates depend on the way the target decoy search strategy is conducted. First, the decoy sequences can be generated either through sequence reversal, shuffling, or randomization. Details on the decoy construction methods are given in the next section. However, various studies have reported that the type of the decoys have little to no effect on the FDR estimates. Second, the tandem spectra can be searched against the combined

target-decoy database (concatenated database) or target database can be searched separately from the decoy database to obtain correct and random score distributions. The concatenated target-decoy database searches are preferred over separate target and decoy database searches because separate searches can produce conservative estimates. This is because in the absence of competition between target and decoy peptides for the same spectra the decoy peptides can receive higher scores relative to the concatenated search strategy.[41, 65, 68] Third, the choice of the formula to compute FDR can produce conservative estimates. The ($2 * N_{decoys} / (N_{decoys} + N_{targets})$) provides conservative FDR estimates relative to the $N_{decoys}/N_{targets}$ formula.

1.9 GENERATION OF DECOY PEPTIDES

A decoy peptide is an in silico generated amino acid sequence that is not present in the original target database (database containing correct peptides for tandem spectra). The database search methods can use a database of decoy peptides to compute the statistical significance value of the peptide-spectrum matches[44] or to determine the score thresholds that separates incorrect from the correct peptide identification and estimation of the corresponding FDR[41]. The FDR estimates and significance values can be estimated from the decoy database based on the assumption that the probability of the incorrect match in the target database can be estimated from the peptide-spectrum matches in the decoy database.[40, 41, 65, 69] Several methods are used to generate the decoy peptides including the sequence reversal method, sequence shuffling method, and random sequence generation method. Each method has its own advantages and disadvantages. A brief description of these methods follows.

The sequence reversal method generates decoy peptides from the original target peptide by sequence reversal (change in the amino-carboxyl orientation) of the target peptide. This method generates the decoy peptides with the same amino acid composition, length and mass distributions as the target peptides. However, the generated peptides are not truly random (i.e., the peptides are generated by reversing the target peptides) and the method cannot be used for palindromic sequences.[68] The sequence reversal method is used in X! Tandem.[45]

The sequence shuffling method generates decoy peptides by randomly shuffling the amino acids in the target peptides. This method preserves the amino acid composition, length and mass distributions of the target peptide. The sequence shuffling method allows repeating the analysis many times by creating different versions of the target peptides than sequence reversal method.[68] This method is implemented in the Crux program to estimate *p-values*.[56]

The random sequence generation method generates decoy peptides by randomly selecting amino acids according to the amino acid frequency and peptide length distributions in the target database. This is undesirable because simple random method cannot preserve amino acid homologies of the target database in the corresponding decoy databases. A better model for the generation of random peptides consists in using a Markov chain model parameterized with the amino acid frequencies of the target database. This approach generates similar amino acid patterns to the target peptides such as acidic or basic regions. The random generation method is implemented in Mascot.[68] The random generating methods can generate many more decoy peptides than the target peptides and this can be undesirable for generating false positive estimation as the relative proportion of decoy to peptide peptides can add decoy bias. Alternatively, the large number of decoy peptides can be used to estimate the significance values for the peptide-spectrum

matches. Prior studies have shown that choice of the type of decoy database have low impact on the accuracy of FDR estimates.[65]

The significance values for the database search programs can be computed either using the parametric approach, semi-parametric approach, or non-parametric approach. The parametric approaches assumes that scores of peptide-spectrum matches follow a certain distribution and the required distribution parameters are obtained from all the matches of particular a spectrum in the sequence database. For example, X! Tandem and OMSSA use the hypergeometric and Poisson distributions to calculate the *E-values*.[43, 45] A semi-supervised parametric procedure uses decoy peptide-spectrum scores to fit a parametric distribution. An example is Crux that calculates *p-values* by fitting Weibull distribution from the target and decoy XCorr scores.[56] The significance values can be calculated using scores of decoy peptides in a non-parametric fashion (without assuming any distribution).[70]

1.10 PERMUTATION TEST

The strengths and limitations of the standard database search programs to identify peptide have been discussed in various comparative studies.[16, 20, 66] The ability of the database search programs to discriminate between the correct and incorrect peptides identifications with good statistical significance values remains an open question. The database search programs must calculate significance of matches irrespective of the peptide size, spectra quality issues such as incomplete fragmentation, low signal to noise ratios, and precursor charge states.[16] The statistical significance values for the peptide identification can be calculated either using parametric, semi-supervised parametric, or non-parametric approaches.

The permutation test or randomization test is a non-parametric statistical significance test to estimate the significance values without assuming any particular distribution for the given data.[71] This makes permutation test useful because in many cases the distribution of the test statistic is usually unknown.[72] Permutation tests can be used for any test statistic or indicator regardless of the original distribution of the test statistic. This makes the permutation test an ideal choice to perform sufficiency analysis and determine the statistic or indicator providing more accurate acceptance or rejection of the null hypothesis. In order to get *p*-values, a null distribution of the test statistic of interest is fitted by calculating all possibilities of the data points through rearrangement (peptide sequences in this case). The *p-value* is calculated as the proportion of rearranged or random peptides receiving equal or better score than the original target peptide. The permutation tests can be categorized into two categories: exact permutation test and Monte Carlo permutation test.

An exact permutation test for a peptide sequence of a given length (L) involves calculation of the test statistic on all possible peptide sequences by sampling amino acids with replacement from a list of 20 standard amino acids. In practice, the enumeration of all possible peptide sequences for the peptides greater than eight amino acids in length would produce a large number of possible permuted peptides. For example, there are $(20)^{10}$ = 10,240,000,000,000 possible ways to generate peptides of ten amino acids in length. These peptides provide possible values of test statistic, the associated distribution and exact *p*-value. The exact test provides an exact *p-value* (the *p-value* observed from an actual experiment) by dividing the number of permutations with score t(r) equal or higher than the original peptides t(s) with total number of permutations (N).

$$P(perm) = \frac{\sum_{i=1}^{N} t(r) \geq t(s)}{N}$$

However, the computation of *p-values* using exact test through enumeration of all possible peptide sequences is not feasible computationally. The second category of permutation test termed as approximate permutation test or Monte Carlo permutation test or randomized permutation test, generates sampling distribution without exhaustively enumerating all possible values of the test statistic.[71, 73] This procedure provides empirical *p*-values that approach there exact *p*-values as the number of sampled permutation increases. The significance levels of the computed *p*-values depend on the number of sampled permutations. For example, to get a *p-value* of 10^{-6} about >=1,000,000 permutation values must be generated. The *p-values* for the Monte Carlo permutation approach are computed in a manner similar to the exact permutation test.

The permutation tests have been extensively used in many bioinformatics areas that include analysis of gene expression data, QTL detection, allelic association analysis, and modeling ChIP sequencing.[73] Likewise, permutation tests can be applied for MS-based peptidomics studies as peptides and proteins are made up of finite (20 standard) number of amino acids. However, it is not practically feasible to enumerate all possible peptide sequences for peptides greater than eight amino acids in length, for example, there are $(20)^{10}$ = 10,240,000,000,000 possible ways to generate different peptides of ten amino acids in length. Therefore, in this entire thesis we used the Monte Carlo permutation approach to convert scores produced by the database search programs into *p-values*.

1.11 FIGURES

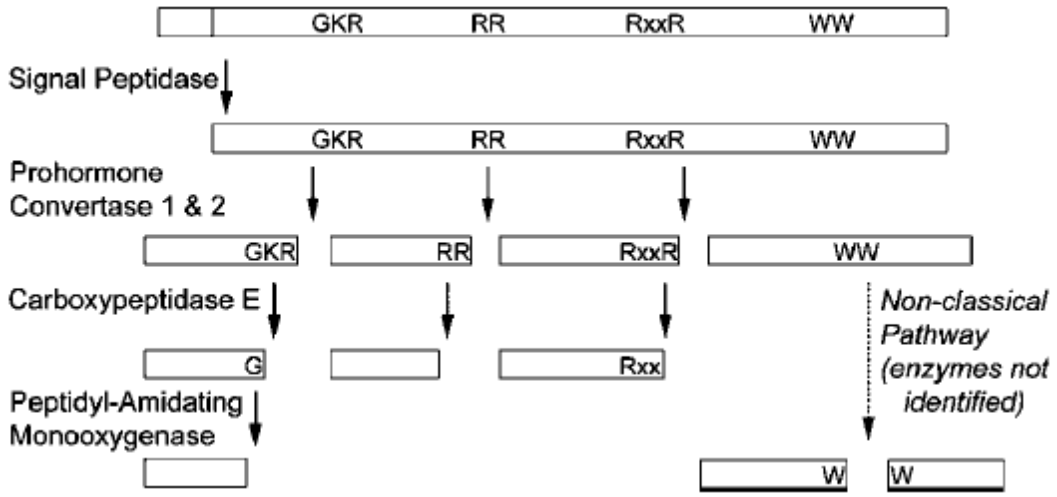

Figure 1.1. Classical and nonclassical neuropeptide processing scheme. First, the N-terminal sequence that drives translocation of the protein into the lumen of the endoplasmic reticulum is co-translationally removed by a signal peptidase. Then, in the classical scheme, the prohormone is typically processed at sites containing Lys-Arg (KR), Arg-Arg (RR), or Arg-Xaa n -Arg, where n is 2, 4, or 6 (RxxR shown in figure). Processing at these basic amino acids involves endopeptidase action by an enzyme such as prohormone convertase 1 or 2 followed by the removal of the C-terminal basic residue(s) primarily by carboxypeptidases E, although an additional enzyme (carboxypeptidase D) is also able to contribute to processing. An amidating enzyme that is broadly expressed in the neuroendocrine system converts C-terminal Gly residues into a C-terminal amide. In addition to this classical pathway, a large number of peptides have been found that result from cleavage at nonbasic residues. An example of this nonclassical pathway for the generation of a peptide previously found in brain is indicated; this fragment of chromogranin B involves cleavage between 2 adjacent Trp residues (WW). Many other nonbasic cleavage sites have been reported, including other hydrophobic residues, short chain aliphatic residues, and acidic residues. The enzymes responsible for the nonclassical pathway are not clear. Some of these nonclassical processing events may occur after secretion and be mediated by extracellular peptidases, although some of the nonbasic mediated cleavages appear to occur within the secretory pathway.[1]

[1] The AAPS Journal, 2, 2005, E449-E455, Neuropeptide-processing enzymes: Applications for drug discovery, Fricker L. D.; with kind permission from Springer.

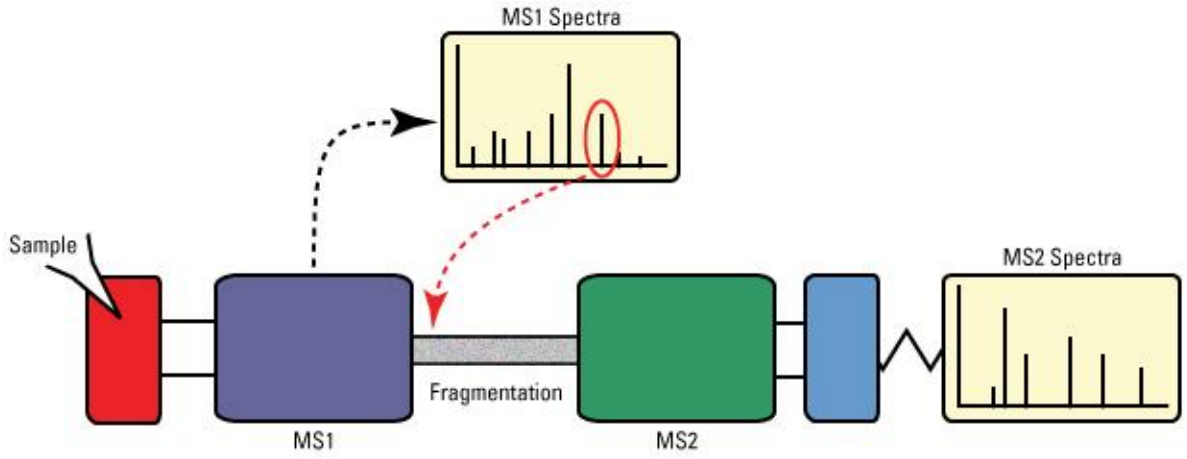

Figure 1.2. Tandem mass spectrometry (MS/MS). A sample is injected into the mass spectrometer, ionized and accelerated and then analyzed by mass spectrometry (MS1). Ions from the MS1 spectra are then selectively fragmented and analyzed by mass spectrometry (MS2) to give the spectra for the ion fragments.While the diagram indicates separate mass analyzers (MS1 and MS2), some instruments can utilize a single mass analyzer for both rounds of MS.[2]

[2] http://www.piercenet.com/method/overview-mass-spectrometry

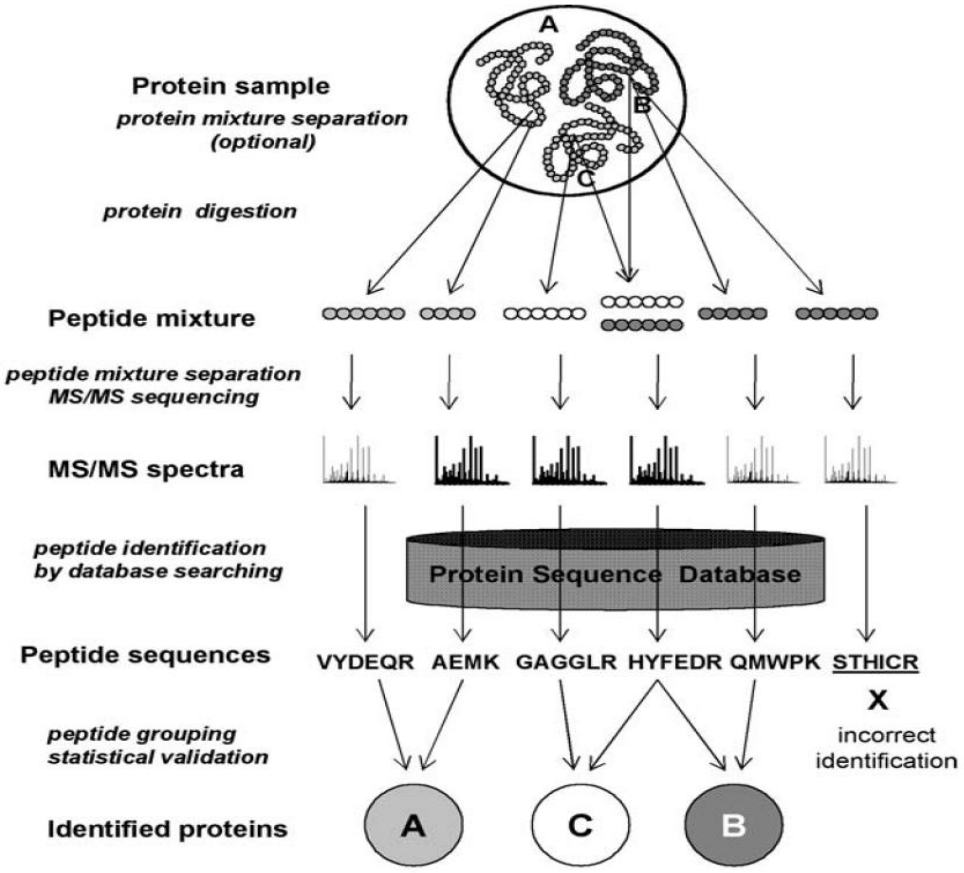

Figure 1.3. General view of the experimental steps and flow of the data in shotgun proteomics analysis. Sample proteins are first proteolytically cleaved into peptides. After separation using one- or multidimensional chromatography, peptides are ionized and selected ions are fragmented to produce signature tandem mass spectrometry (MS/MS) spectra. Peptides are identified from MS/MS spectra using automated database search programs. Peptide assignments are then statistically validated and incorrect identifications filtered out (peptide STHICR). Sequences of the identified peptides are used to infer which proteins are present in the original sample. Some peptides are present in more than one protein (peptide HYFEDR), which can complicate the protein inference process.[3]

[3] Springer and the Methods in Molecular Biology, 367, 2007, 87-119, Protein identification by tandem mass spectrometry and sequence database searching, Nesvizhskii A. I.; with kind permission from Springer.

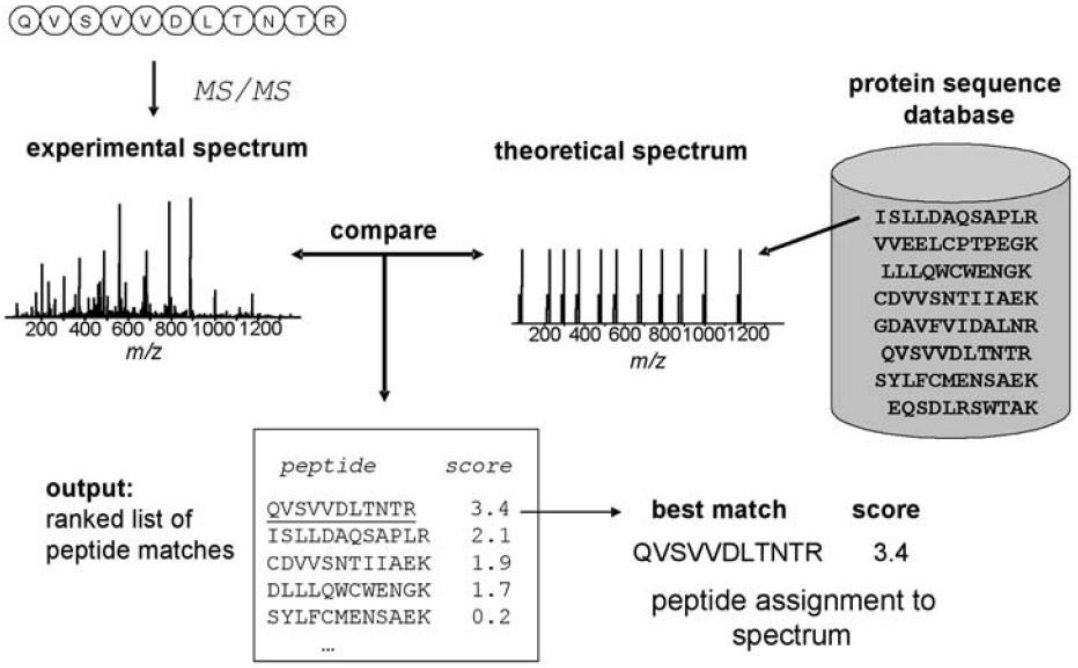

Figure 1.4. Tandem mass spectrometry (MS/MS) database searching. Acquired MS/MS spectra are correlated against theoretical spectra constructed for each database peptide that satisfies a certain set of database search parameters specified by the user. A scoring scheme is used to measure the degree of similarity between the spectra. Candidate peptides are ranked according to the computed score, and the highest scoring peptide sequence (best match) is selected for further analysis.[4]

[4] Nature Methods, 4, 2007, 787-797, Analysis and validation of proteomic data generated by tandem mass spectrometry, Nesvizhskii *et al.*; with kind permission from Nature Publishing Group.

CHAPTER II: ACCURATE ASSIGNMENT OF SIGNIFICANCE TO NEUROPEPTIDE IDENTIFICATIONS USING MONTE CARLO K-PERMUTED DECOY DATABASES

2.1 NOTES AND ACKNOWLEDGMENTS

The material presented in this chapter is a preprint version of the article "Akhtar, M. N.; Southey, B. R.; Andren, P. E.; Sweedler, J. V.; Rodriguez-Zas, S. L. Accurate assignment of significance to peptide identifications using Monte Carlo k-permuted decoy databases" published in PLOS ONE (2014). This work was completed in Dr. Sandra Rodriguez-Zas Bioinformatics Laboratory at University of Illinois Urbana-Champaign (USA) in collaboration with Dr. Sweedler at University of Illinois Urbana-Champaign (USA) and Dr. Andren at University of Uppsala (Sweden). The work is focused on estimation and evaluation of significance values for the neuropeptide identifications from the database search programs using whole sequence permutation databases. The support of NIH (Grant Numbers: R21 DA027548, P30 DA018310 and R21 MH096030) is greatly appreciated.

2.2 ABSTRACT

In support of accurate neuropeptide identification in mass spectrometry experiments, novel Monte Carlo permutation testing was used to compute significance values. Testing was based on k-permuted decoy databases, where k denotes the number of permutations. These databases were integrated with a range of peptide identification indicators from three popular open-source database search software (OMSSA, Crux, and X! Tandem) to assess the statistical significance of neuropeptide spectra matches. Significance *p-values* were computed as the fraction of the sequences in the database with match indicator value better than or equal to the true target spectra. When applied to a test-bed of all known manually annotated mouse neuropeptides, permutation tests with k-permuted decoy databases identified up to 100% of the neuropeptides at *p-value* < 1×10^{-5}. The permutation test *p-values* using hyperscore (X! Tandem), *E-value* (OMSSA) and Sp score (Crux) match indicators outperformed all other match indicators. The robust performance to detect peptides of the intuitive indicator "number of matched ions between the experimental and theoretical spectra" highlights the importance of considering this indicator when the *p-value* was borderline significant. Our findings suggest permutation decoy databases of size 1×10^{5} are adequate to accurately detect neuropeptides and this can be exploited to increase the speed of the search. The straightforward Monte Carlo permutation testing (comparable to a zero order Markov model) can be easily combined with existing peptide identification software to enable accurate and effective neuropeptide detection. The source code is available at http://stagbeetle.animal.uiuc.edu/pepshop/MSMSpermutationtesting.

2.3 INTRODUCTION

Neuropeptides participate in cell to cell communication and regulate many biological processes such as behavior, learning, and metabolism.[1] Mass spectrometry has revolutionized neuropeptide characterization and quantification.[74-79] However, detection is complicated by the neuropeptide size (typically 3 to 40 amino acids long) and by the complex post-translational processing that includes cleavage, and amino acid modifications of prohormones into neuropeptides.[1, 6]

Database search programs are commonly used to identify peptides from tandem mass spectrometry experiments.[41] These programs generate in silico theoretical spectra from target databases of known peptide sequences that have masses within a range (tolerance) of the observed peptide mass. The in silico spectra are then compared to the observed experimental spectra and indicator scores that signify the closeness of the match are computed. To assess the statistical significance of these matches, the observed-target match indicator is compared to the distribution of indicator values under the null hypothesis of no match using various methods. In the popular target-decoy approach, the experimental spectra are compared to spectra from a decoy database consisting of peptides sequences that were generated by reverting or reshuffling the amino acids in the sequences of the target database.[16, 41, 46]

For neuropeptide identification, the target-decoy approach can result in false negatives because the small size of many neuropeptides leads to low observed-target match indicator values and consequently low significance levels.[16] Furthermore, the small size of many neuropeptide leads to few decoy reshuffled sequences and the resulting granularity of the null distribution of

decoy scores further lowers the significance levels.[16, 20, 58, 80, 81] At the protein level, alternative identification approaches have attempted to address the challenge of assessing statistical significance.[82, 83] However, the implementations of the previous approaches do not work with widely used database search programs, do not use all the information resulting from the mass spectrometry experiment, and are biased by peptide length or assume one-direction progressive processing. Approaches that rely on fewer limiting assumptions and that use all the information available need to be evaluated.

Permutation tests are well-suited for neuropeptide database searches by helping to overcome the finite combination of amino acids from small neuropeptides and do not rely on directional assumptions. Furthermore, permutation testing provides strong control of Type I errors thus minimizing the incidence of false positive results.[84] Under the null hypothesis of no match, the experimental spectrum of a peptide is the result of a random sequence of amino acids provided that the total mass is close to the experimental mass. This requirement stems from the database search program strategy that only accepts sequences within a user determined range of the experimental spectra. The permutation statistical significance under the null hypothesis is then generated by using a decoy database of considering all possible amino acid sequences within the predetermined range of the experimental spectra. Under the null hypothesis any amino acid can be present at any position of the sequence, thus, addressing the exchangeable assumption required by the permutation test.[84]

Monte Carlo sampling is used to reduce the number of possible sequences while providing an unbiased estimate of the *p-value*. Furthermore, the loss in statistical efficiency when estimating the *p-value* decreases with increasing number of random samples.[84] The main advantage of the

Monte Carlo permutation approach proposed over existing decoy generation based on sequence reversion or reshuffling of the target sequence is the improved definition of the null distribution. The larger number of decoy sequences results in lower granularity and, thus, more precise assessment of the statistical significance of the observed matches. Two major advantages of the Monte Carlo permutation approach proposed over existing dynamic programming approaches[82, 83] is the simplicity of integration to existing database search programs, the use of all spectra information available and consideration of all possible spectra matching processes.

This study demonstrates the use Monte Carlo permutation testing to overcome the limitations of current protein identification approaches to accurately assess neuropeptide statistical significance. This approach combines and extends the model-free property of current decoy databases with the more extensive search of dynamic programming approaches. The aims are: (1) to develop permutation resampling methodology that can be easily integrated with existing peptide database search software, and (2) to demonstrate the advantages of this approach to provide accurate measures of neuropeptide match significance using ideal and real experimental neuropeptide spectra. Supporting objectives were: (1) to develop and implement complementary novel permuted databases; (2) to determine the number of permutations required for accurate significance levels; and (3) to identify the neuropeptide match indicators within and across programs that are better suited to provide accurate statistical significance.

2.4 MATERIALS AND METHODS

TANDEM SPECTRAL DATASET AND TARGET DATABASE

Tandem mass spectra from a comprehensive list of 103 experimentally-obtained and manually annotated mouse neuropeptides were obtained from the SwePep database (http://www.swepep.org). These spectra were obtained using linear ion trap mass spectrometer coupled with liquid chromatography and electrospray ionization source.[25] Neuropeptides were manually validated after identification using the X! Tandem database search program.[45] The independent manual annotation step also ensured that the subsequent software comparison would not be biased in favor of the X! Tandem database search program. Of these, 80 neuropeptides were unmodified and the remaining 23 encompassed post-translational modifications (PTMs) including C-terminal amidation, N-terminal acetylation, phosphorylation, pyro-glutamination and oxidation. The spectra corresponded to 5, 68, 25, and 5 peptides that had precursor charge states +1, +2, +3 and +4, respectively, and all charge states were observed in modified and unmodified peptides.

Ideal uniform spectra of all possible *b*- and *y*-ions with +1 product charge state were simulated for 103 annotated experimental spectra. The ideal spectra also included all the PTMs identified in the corresponding experimental spectra. The neutral mass loss peaks due to loss of single water or ammonia molecules from the *b*- and *y*-ions were simulated regardless of their position in the ions sequence.

A comprehensive target database of 618 mouse neuropeptides was obtained from the PepShop database[32] (http://stagbeetle.animal.uiuc.edu/pepshop). This target database encompassed

the neuropeptides corresponding to the 103 tandem spectra studied. The neuropeptides in the PepShop were assembled from the known 95 mouse prohormones present in SwePep[25] and UniProt[24] complemented with NeuroPred[85] predictions. The neuropeptides in the target database ranged from 2 to 223 amino acids in length because this included all known experimentally confirmed mouse neuropeptides as well as all possible intermediates and other peptides produced during the processing of prohormones. The target database of neuropeptides is available at http://stagbeetle.animal.uiuc.edu/pepshop/MSMSpermutationtesting.

DATABASE SEARCH PROGRAMS AND DATABASE SEARCHING

Three open source database search programs were used in this study: Crux (version 1.37),[44] OMSSA (version 2.1.8),[43] and X! Tandem (version 2013.02.01.1).[45] These commonly used open source programs were selected because the code could be modified to ensure comparable search parameter specification and enabled to retrieve intermediate indicators of the strength of the match between the observed and target or decoy spectra. The observed-target or observed-decoy spectra match indicators extracted from OMSSA were: number of matched fragment ions, lambda or Poisson mean match indicator, Poisson probability of the lambda match indicator, and corresponding *E-value* of the match (Poisson probability multiplied by the effective database size). The spectra match indicators extracted from X! Tandem were: number of matched fragment ions, intermediate convolution score (product of the intensities of the shared *b*- and *y*-fragment ions between experimental and theoretical spectra), hyperscore (factorial of the number of matching *b*- and *y*-ions multiplied by the convolution score), and *E-value* (calculated from the distribution of hyperscores scores). The spectra match indicators extracted from Crux were: number of matched

fragment ions, Sequest Sp score (Sp), cross-correlation score (XCorr), deltaCn score (ΔCn) and *p-value* that is calculated from the Weibull distribution fitted to the XCorr scores of observed-theoretical spectra matches.[56]

For comparable neuropeptide identification across the three programs the following search parameters from our prior research[16] were used: (1) precursor ion tolerance: 1.5 Da; (2) fragment ion tolerance: 0.3 Da (OMSSA and X! Tandem); mz-bin-width: 0.3 (Crux) (3) searches were performed with and without PTMs. The PTMs evaluated were: amidation, phosphorylation, N-terminal acetylation, acetylation of lysine, pyroglutamination of glutamine, methylation of lysine and arginine residues, sulfation of tyrosine residue, and oxidation of methionine; (4) "protein" (OMSSA) or "enzyme: custom cleavage site" (X! Tandem and Crux) to prevent peptide cleavage since the detection of neuropeptides does not involve protease digestion; (5) fragment ion charge: default values; (6) OMSSA "ht" option was set to eight to filter database peptides that have at-least one theoretical fragment ion match to one of the top eight most intense peaks in the observed spectra; and (7) peptide mass: monoisotopic; 8) Crux *p-values* were computed using 1000 Weibull points because this information provides more accurate *p-values* than the default 40 Weibull points.[16]

PERMUTATION APPROACH AND K-PERMUTED DECOY DATABASES

A Monte Carlo permutation test approach based on biological, computational and statistical considerations was used to generate decoy sequence databases that, in turn, can be used by all database search programs without the need to modify the original program code. Applying the

same strategy used by the database search programs, candidate mouse neuropeptides within 12 Da of the precursor mass of the 103 studied neuropeptides were considered for permutation. This resulted in 236 peptide sequences available for permutation. The 12 Da threshold enabled the creation of a single catalog of peptides independent of charge state tolerance because this databases included all peptides within a 3 *m/z* ion mass tolerance of the target peptide at a charge state of +4. This single catalog was used to create target and was the basis to generate all the decoy permutation databases evaluated. **Figure 2.1** depicts the correspondence between the lengths of neuropeptides in the target database, the 103 experimental neuropeptides and the neuropeptides that fall within 12 Da of the 103 peptides. Decoy peptide sequences were randomly generated by sampling the 19 amino acids from the candidate peptide list (leucine and isoleucine were considered the same amino acid due to the high similarity of the neutral masses). The resulting libraries are comparable to those generated from a Markov model of order zero. A permutated database of only 10-amino acid long peptides would lead to 6.13×10^{12} permuted sequences. Due to this potential size of a database encompassing all possible permutations, a Monte Carlo permutation approach was used to generate a random sample of all possible sequences. These permuted sequences were collected into a single database after removal of duplicate peptides and sequences present in the target database. This procedure was used to generate k-permuted decoy sequence databases and the numbers of unique permuted sequences per candidate peptide (k) were: 10^3 ($K10^3$ with 236,000 decoy peptide sequences), 10^4 ($K10^4$ with 2,360,000 decoy peptide sequences), 10^5 ($K10^5$ with 23,600,000 decoy peptide sequences), and 10^6 ($K10^6$ with 236,000,000 decoy peptide sequences). The target database was appended to each of the four k-permuted databases to create a combined target-k-permuted decoy database. The combined database search

is more accurate than separate database searches and to avoid zero *p-value*.[65, 69, 73, 84] This strategy also removed potential database size dependency of the match indicators between target and permuted sequences because the correct match was evaluated under the same database sizes as the permuted databases.

The search of spectra against the k-permuted decoy databases produced many matches that were indistinguishable from each other based on the indicators reported by the programs (e.g., number of matched ions, hyperscore, convolution score, and *E-value* for the X! Tandem). Matches were considered "homeometric"[86] when the matches had the same indicator values across programs and the matched peptides masses were within ±1.5 Da from each other. Matches were considered "heterometric" when the matches differed in at least one indicator value or the matched peptides masses differed by more than ±1.5 Da from each other. **Figure 2.2** depicts the number of peptides with homeometric matches ranging from 1 to 10 for the $K10^6$ k-permuted decoy database across the three databases search programs. Homeometric matches were counted only once while calculating the number of random peptides that have an indicator value equal or better than the true target peptide. This strategy resolved the challenge that database search programs were not able to differentiate between such matches that are technically redundant and ensured the calculation of permutation *p-values* that were unbiased by these effects.

For each database search program and target sequence, the observed tandem spectra were searched for matches within each combined target-k-permuted decoy spectra. The permutation *p-values* were estimated as the fraction of combined target-k-permuted decoy peptides, excluding any homeometric matches that have a matching indicator score equal or better than the score of target peptide.

A comprehensive evaluation of the k-permuted decoy approaches, programs, and peptide match indicators was undertaken including: (a) Search for ideal uniform simulated spectra against the target database using all three database search programs; (b) Search for real tandem spectra against the target database using all three database search programs; (c) Search for the 80 tandem spectra containing no PTMs against the $K10^3$, $K10^4$, $K10^5$, and $K10^6$ target-k-permuted decoy databases without PTM specification using all three database search programs; (d) Search for the 80 tandem spectra containing no PTMs against the $K10^5$ k-permuted database with PTM specification using all three database search programs; and (e) Search for the 23 tandem spectra containing PTMs against the $K10^5$ k-permuted database with PTM specification using OMSSA and X! Tandem. Crux was excluded from this last comparison due to considerable amount of search time required.

2.5 RESULTS AND DISCUSSION

Results from a three step benchmarking strategy were used to evaluate the performance to detect neuropeptides using target-k-permuted decoy databases. First, a baseline performance was obtained by comparing ideal simulated spectra against a standard "target database" using the three database search programs. Then, observed tandem spectra were matched to a target database. Lastly, the observed tandem spectra were matched to different target-k-permuted decoy databases. The source code to generate k-permuted decoy databases is available at http://stagbeetle.animal.uiuc.edu/pepshop/MSMSpermutationtesting.

Table 2.1 summarizes the results from the three database search programs when 103 ideal uniform spectra were simulated with all *b*- and *y*-ions including neutral mass losses and searched against the target database. The search of ideal simulated spectra demonstrated the ability of the database search methods to assign *E*- or *p-values* to each peptide-spectrum match in the absence of technical or biological noise.[16]

The three programs matched all unmodified peptides correctly at *E*- or *p-value* < 2 x 10^{-1}. At *E*- or *p-value* < 1 x10^{-2}, OMSSA, X! Tandem, and Crux identified 80 (100%), 80 (100%), and 73 (91.25%) peptides, respectively. This trend was consistent with previous study that compared Crux, OMSSA and X! Tandem.[16] Our study confirmed the lower significance values that Crux computes for peptides less than 45 amino acids in length.[16] OMSSA *E-values* averaged more significant matches than X! Tandem for the 32 peptides that were less than 13 amino acids in length. However, for the 48 peptides longer than 12 amino acids in length, the difference in significance levels of X! Tandem and OMSSA decreased on the average with 8, 18, and 22 peptides getting lower, equal, and better significance levels for the X! Tandem than OMSSA, respectively.

For the 23 neuropeptides with PTMs and an *E*- or *p-value* < 1 x 10^{-1}, OMSSA, X! Tandem, and Crux correctly detected 23 (100%), 18 (78.26%), and 23 (100%) peptides, respectively. X! Tandem failed to correctly match five peptides with N-terminal acetylation modification instead these five peptides were matched with incorrect internal acetylation modification at 9^{th} lysine residue. The failure in the peptide detection of X! Tandem was only observed when multiple PTMs

were specified in the search specification. The five peptides were correctly detected when only N-terminal acetylation was used in the search specification. At *E*- or *p-value* < 1 x 10^{-2}, 23 (100%), 18 (78.26%), and 22 (95.66%) peptides were detected by OMSSA, X! Tandem, and Crux, respectively. The three peptides that were not significant for OMSSA at *E-value* < 1 x 10^{-4} all had a pyroglutamination (Q residue) modification. Two of these peptides, somatostatin [87-100] (QRSANSNPAMAPRE; charge state +2) and secretogranin-2 [205-216] (QELGKLTGPSNQ; charge state +1), were significant for the X! Tandem and Crux at *E*- or *p-value* < 1 x 10^{-4}. A nine amino acid long peptide secretogranin-1 [667-675] (QKIAEKFSQ; charge state +2) was not significant for all three programs at *E*- or *p-value* < 1 x 10^{-4}, while the same peptide was missed by the Crux at *p-value* < 1 x 10^{-2}.

PEPTIDE DETECTION USING OBSERVED SPECTRA AND A TARGET DATABASE

Table 2.2 summarizes the performance of the three database search programs when the 80 experimental tandem spectra containing no PTMs were searched against the target database. All peptide assignments by the three database search methods were correct at *E*- or *p-value* < 5 x 10^{-1}. At *E*- or *p-value* < 1 x 10^{-2}, OMSSA, X! Tandem and Crux detected 80 (100%), 71 (88.75%), and 63 (78.75%) peptides, respectively. The higher number of significant peptide detections by OMSSA relative to Crux was consistent with the prior reports.[16] The three search methods were less accurate on 23 observed spectra with PTMs when searched against the standard target database (**Table 2.2**). From the correctly matched peptides for each program, at *E*- or *p-value* < 1 x 10^{-2}, OMSSA, X! Tandem and Crux detected 20 (86.95%), 15 (65.21%), and 17 (73.91%) peptides, respectively.

The 80 spectra without PTMs were searched against the target database using three database search programs and with PTM specifications. X! Tandem peptide detection significance levels for the 76, 3, and 1 target peptide remained unchanged, decreased, and increased, respectively, relative to the searches involving no PTMs. The changes in the significance levels of the four peptides were due to higher number of candidate peptides available in the PTM searches which in turn changed the estimation parameters used in the *E-value* computation. The OMSSA peptide detection significance levels decreased for the majority of the previous peptides (75 out of 80 peptides) or remained unchanged (5 out of 80 peptides) when searches included PTMs, respectively. Crux peptide detection significance levels were improved when searches included PTMs with 65 and 29 peptide detections at *p-value* $< 1 \times 10^{-2}$ and $< 1 \times 10^{-4}$, respectively. Comparison of peptide detections across PTM scenarios indicated that at *p-value* $< 1 \times 10^{-2}$, 54 peptides were detected by both scenarios, 11 peptides were detected in the PTM scenario, 9 peptides were detected in the no PTMs scenario, and 6 peptides were not detected by either scenario. The target peptides with low XCorr scores remained undetected either across both scenarios or with PTM search. The clear positive correlation between significance level and XCorr score for the PTM searches relative to the searches without PTMs could be due to the higher number of low scoring matches in the searches with PTMs than without PTMs. The Crux resampling from the low scoring matches might have resulted in a shift on the distribution of XCorr scores towards lower scores than the target peptides XCorr scores.

Table 2.3 summarizes the $\log_{10}$ transformed of the *E-values* to the target database and permutation *p-values* computed for the X! Tandem indicators: number of matched ions, hyperscore, *E-value*, and convolution score using the 80 spectra without PTMs across the four target-k-permuted decoy databases studied. The permutation *p-values* from number of matched ions, hyperscore and *E-value* showed that the X! Tandem *E-values* from the target database were dramatically underestimated (less significant) for most target peptides. Detection and significance level using the number of ions matched, hyperscore and *E-value* were almost the same across all target-k-permuted decoy databases. Only at the 10^6 permutations did the *p-values* for number of ions matched started to differ from the *p-values* from the hyperscore and *E-value* match indicators. This trend was expected as the hyperscore is a function of the product of factorial of the number of matched ions and the ion intensity values and *E-value* is a function of the hyperscore.

The convolution score resulted in fewer target peptide identifications with higher number of sequence permutations due to relative increase in the number of decoy matches with equal or better scores. From the $K10^3$, $K10^4$, $K10^5$, and $K10^6$ target-k-permuted decoy databases, 72 (90%), 31 (39%), 9 (11 %), and 10 (13%) peptides were identified at *p-value* $< 1 \times 10^{-2}$, $< 1 \times 10^{-3}$, $< 1 \times 10^{-4}$, and $< 1 \times 10^{-4}$, respectively. These results showed that the convolution score alone was less suitable to discriminate between true target and decoy matches than the hyperscore and *E-value*.

Comparison of the *p-values* obtained from the target-k-permuted decoy number of matched ions, hyperscores and convolution scores suggested that roughly 10^5 permutations were required for significant *p-value* computations using the convolution scores. Higher number of sequence

permutations provided better separation between the significance levels of the three indicators. There were 7 peptides with *E-values* < 10^{-7} from the target database indicating that the lower bound of *p-values* appeared to be far smaller than the limit provided by the $K10^6$ permuted database. Comparable performance (significance level) using number of matched ions and hyperscore were observed with fewer permutations or lower significance thresholds. This novel finding suggests that more significant detections can be obtained by permuting the X! Tandem hyperscore and number of matched ions indicators, even with a relatively small k-permuted decoy database size.

CRUX PEPTIDE IDENTIFICATION USING A K-PERMUTED DECOY DATABASE

Table 2.4 summarizes the $\log_{10}$ transformed permutation *p-values* computed for the Crux match indicators: number of matched ions, XCorr, ΔCn, and Sp using the 80 spectra without PTMs across the four target-k-permuted decoy databases. Higher number of sequence permutations increased the significance values using the number of matched ions and Sp. This trend was due to the lower number of matched ions and Sp scores of the decoy peptide matches relative to the target peptides. The two non-detected peptides could be attributed to the low number of decoy candidates for those peptides rather than to an increase in the number of decoy peptides with equal or better scores. The hindering effect on the match significance of better or equal decoy matches on Sp was more evident with the large decoy databases at *p-value* < 1×10^{-5}.

Peptide detection was less significant when using XCorr relative to Sp and number of matching ions. The drop in significance level with increase in threshold and database size was due

to the higher number of decoy peptides reaching XCorr levels better or equal than the target peptides. The detection and significance computation using XCorr and ΔCn (the difference in XCorr between candidates) was similar across all target-k-permuted databases which reflects that the range of these match indicators stabilized. The range of possible XCorr values was limited by the number of observed spectrum peaks because the background adjustment is expected to be constant across permuted database sizes. This result indicates that only a relatively few permuted sequences are required to cover the range of XCorr values and that higher number of permutations offer greater precision to detect match differences.

OMSSA PEPTIDE IDENTIFICATION USING A K-PERMUTED DECOY DATABASE

Table 2.5 summarizes the $\log_{10}$ transformed permutation *p-values* calculated for the OMSSA match indicators: number of matched ions, lambda match indicator, *p-value*, and *E-value* using the 80 spectra without PTMs across the target-k-permuted decoy databases. Comparison between the target database and the permutation *p-values* indicated that most peptides were accurately estimated by OMSSA suggesting that the k-permuted database size was unimportant. Examination of the few peptides with underestimated *E-values* suggested that these peptides had fewer intense MS/MS ion peaks resulting in lower 75% quartile values than peptides of similar size with lower *E-values*. This result indicates that OMSSA *E-values* may be less reliable in the presence of multiple low intensity spectra peaks.

Detection and significance computation using the number of matched ions, OMSSA *p-value* and *E-value* indicators was identical across all k-permuted decoy databases. However, the

lambda parameter was less suitable than the other OMSSA match indicator to discriminate matches than the other match indicators. Differences in the lambda indicator for the same observed spectrum were mainly determined by the total number of theoretical *m/z* values for product ions and hence by the length of the decoy peptide sequence. After a relatively few permutations, the range of possible sequences is determined such that fewer permutations are required to determine the distribution of the lambda parameter than other match indicators.

IMPACT OF PTM ON PEPTIDE IDENTIFICATION USING A K-PERMUTED DECOY DATABASE

Searches of 80 peptides with no PTMs including the specification of common neuropeptide PTMs improved the significance of the detection in target-k-permuted decoy databases. Using X! Tandem, all 80 observed peptides were identified at *p-value* $< 1 \times 10^{-5}$ using the number of matched ions and hyperscore indicators in the $K10^5$ permuted database, while convolution score indicator detected only 7 (8.75%) peptides. Consistent with searches without PTMs using the OMSSA program, when the searches included PTMs the number of matched ions and *E-value* indicators provided more significant permutation *p-values* than the lambda indicator. For Crux, specification of PTMs reduced the performance (significance levels) of the number of matched ions, XCorr, and Sp indicators in the $K10^5$ database. The lower significances was due to corresponding increase in the decoy peptides with equal or better scores than the target peptides with increase in decoy database size when PTMs are considered in the search. Using the $K10^5$ permuted database, OMSSA and X! Tandem correctly identified the 20 and 17 of spectrum with PTMs as the first match, respectively. Both programs correctly identified the same 16 peptides, 6 peptides were identified by only one program and 1 peptide was not detected by either program.

There were 4 peptides unmatched by X! Tandem only and the unmodified forms were matched outside the top 20 matches. The unmatched peptide, acetyl-YGGFMTSEKSQTPLVT, was undetected by OMSSA both in the target or k-permuted databases. X! Tandem was able to match the correct sequence, however the match has an additional amidation. Manual evaluation would have corrected the match as the amidation was on an unexpected amino acid and the non-amidated form was closer to the precursor mass then the amidated form.

The remaining 2 peptides that were unmatched by OMSSA were both amidated. One peptide, SYSMEHFRWGKPV-amide, was correctly identified as the 15th best match by OMSSA with the unamidated form providing the best match. The difference in monoisotopic mass between modified and unmodified was less than 1 Da. The experimental spectrum had a precursor m/z value of 541.70 with an assigned a 3+ charge state. At a 3+ charge state the predicted m/z values were 541.9294 and 541.6014 for the unmodified form and amidated forms, respectively. Biologically the unmodified form would be identified as a probable match since this sequence is an intermediate in the amidation process and the unmodified sequence is uncommon among neuropeptides because this form lacks the terminal G-residue after cleavage.[87] Consequently this unmodified peptide could be considered a match for OMSSA.

COMPARISON OF PEPTIDE DATABASE SEARCH PROGRAMS

Overall the k-permuted decoy databases allowed the detection of more peptides based on real spectra than the use of the standard target database regardless of the database search program. The search of spectra against the k-permuted decoy databases produced many matches that were

indistinguishable from each other based on the indicators reported by the programs (e.g., number of matched ions, hyperscore, convolution score, and *E-value* for the X! Tandem). Permutation testing is computational demanding even with Monte Carlo sampling (**Table 2.6**). The increase in time across permutated database sizes is a consequence of the exponential increase in the number of sequences evaluated. However, the $K10^5$ database provided adequate results and all programs completed the search within 35 CPU minutes using a single process Intel® Core™ i7-3770 CPU @ 3.40GHz. This timing is the result of single-processor searches that ignored possible parallel processing of individual spectra. The advantages of Monte Carlo permutation approaches to assess the statistical significance of neuropeptide matches could be further advanced by simultaneously running groups of observed spectra using parallel processing.

An alternative approach to generate a permutated database is to perform targeted permutation of specific regions such as the terminal amino acids to disrupt *b*- and *y*-ion series. While other regions can be permuted, the advantage of permuting only the terminal peptides is that this strategy is independent of peptide size. The size of the required database quickly increases from 84,960 sequences per target peptide when one terminal position was permuted to 47,045,880 sequences per target peptide when three terminal positions were permuted. Evaluation of terminal permuted databases demonstrated that this approach offered similar yet less significant matches than the whole sequence permuted database approach. Also, this permutation approach had the disadvantage of providing a large number of homeometric matches since experimental ions near the termini are required to differentiate the order of amino acids. Thus, results from this approach are not reported.

With the goal of accurate significance evaluation of protein matches, dynamic programming-related approaches have been proposed.[82, 83] However, dynamic programming assumes that a problem (i.e., spectra matching) can be divided into independent components. In the context of tandem spectra, any division based on sequence location creates dependent components because changing an amino acid in any location will change both the *b*- and *y*-ion fragment series. Further any mass change must be balanced by a corresponding change in another part of the sequence such that the overall mass is within the specified tolerance of the original mass. Also, the implementation of these approaches limit high computational requirements by limiting the information considered or through analytical assumptions. These strategies resulted in non-exhaustive libraries that could lead to biased statistical significance assessment. In one case, the algorithm used is location based such that the only one ion series can be used[83] due to interrelationship between ion series and that precursor must remain within the preset tolerances. However, using only one series is not as effective as using both ion series and that one ion series can be more informative than the other series.[16] In the other case, the score for a given number of matched peaks is assumed to encompass the score from fewer matched peaks.[82] This assumption fails when different sets of peaks are being matched from the same peptide and the number of peaks in common changes. Both strategies do not consider the optimal starting location such that a peptide will be dropped from consideration when a region of the spectrum has a poor match score despite the higher score in other unevaluated regions. The published algorithms appear to lack error corrections for common problems of incorrect peak assigned due to charge state, presence of chimeric peptides, and missing peaks. Also, both dynamic programming strategies do not have a clear approach to account for peptide length that has been proven to bias the statistical significance

of neuropeptides identifications.[16] Lastly, both approaches cannot be directly applied to the open source X! Tandem, Crux and OMSSA unlike the straightforward permutation approach proposed in this study. Although the lack of comparable basis challenges the benchmarking of strategies, the Monte Carlo permuted database approach proposed addresses the previous limitations while enabling simple integration to database search programs and prompt results.

2.6 CONCLUSIONS

The present study demonstrated that the k-permuted decoy database is an effective and computationally feasible approach to accurately calculate the statistics of neuropeptide matches from complex tandem MS datasets. Unlike other proposed methods to control multiple testing, such as target-decoy approaches, permutation testing provided strong control of Type I errors such that neuropeptides are detected at high confidence of significance. The implication of this finding is that an extensive decoy database is not required to accurately detect neuropeptides and this can be exploited to increase the speed of the search.

This study demonstrated the relative superiority of specific detection indicators for database search programs. The indicators *E-value*, hyperscore, and Sp score from the OMSSA, X! Tandem, and Crux programs, respectively, performed better than other indicators. The results indicated that 10^5 permutations per peptide were sufficient to provide significant peptide identifications. Indication of the suitability of the Monte Carlo permutation approach using 10^5 permutations was the capability of all three database search programs to detect all or nearly all neuropeptides at *p-value* < 10^{-4} and the absence of a trend for lower statistical significance with higher permutation number. A promising finding is the robust performance of the simple indicator,

number of matched ions between the experimental and theoretical spectra to detect peptides. This intuitive indicator identified the vast majority of the peptides also identified by other indicators such as hyperscore, Sp and *E-value* that rely on assumptions or parametric specifications. This result also highlights the importance of considering the number of matched ions when a match is borderline significant. The results have shown that, in conjunction with database search programs, the k-permuted sequence databases allowed the detection of more peptides and exhibited high consensus among the various indicators and database search programs.

The permutation testing approached developed here can easily be integrated into standard database search programs to compute spectrum specific *p-values* for any indicator reported by the program. Through the generation of decoy peptides, the permutation approach could offer insights into unknown or unexpected neuropeptides (including those resulting from PTMs or polymorphisms or chimeras) not present in the target database. Further, the k-permuted databases can be generated once and shared between programs and the community.

2.7 FIGURES

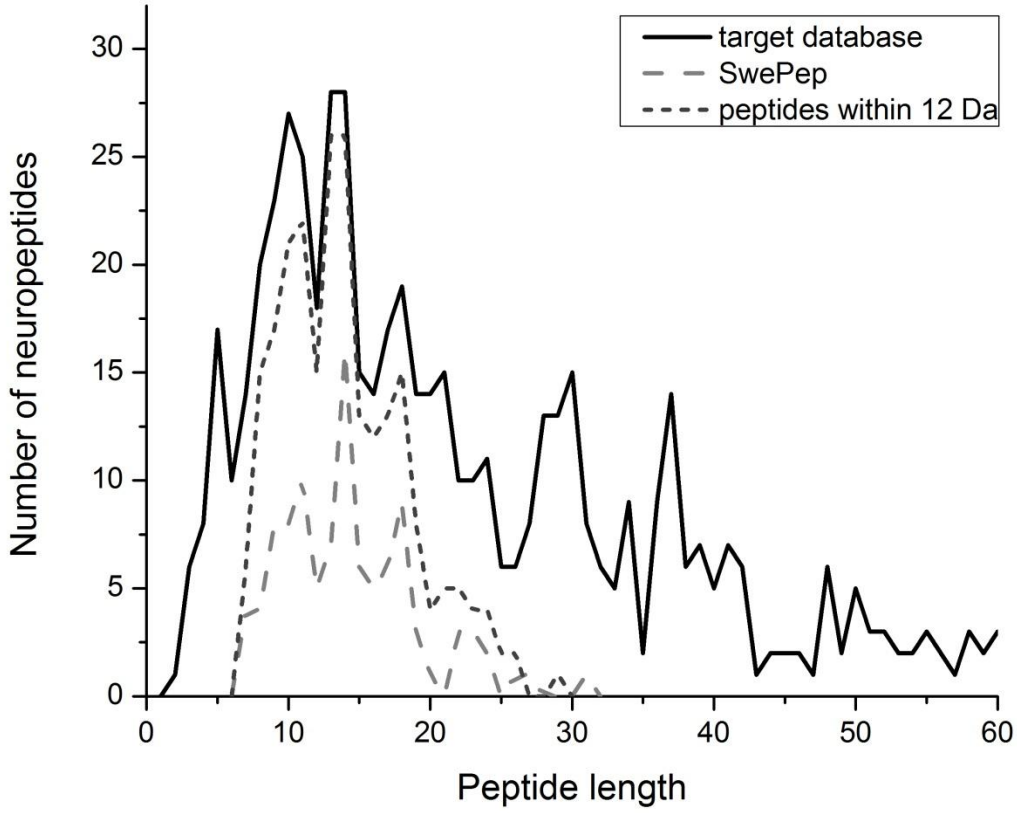

Figure 2.1. Distribution of neuropeptides length in target database peptides less than 60 amino acid in length are shown, 103 MS/MS peptides, and 236 peptides that fall within ±12 Da of the SwePep peptides.

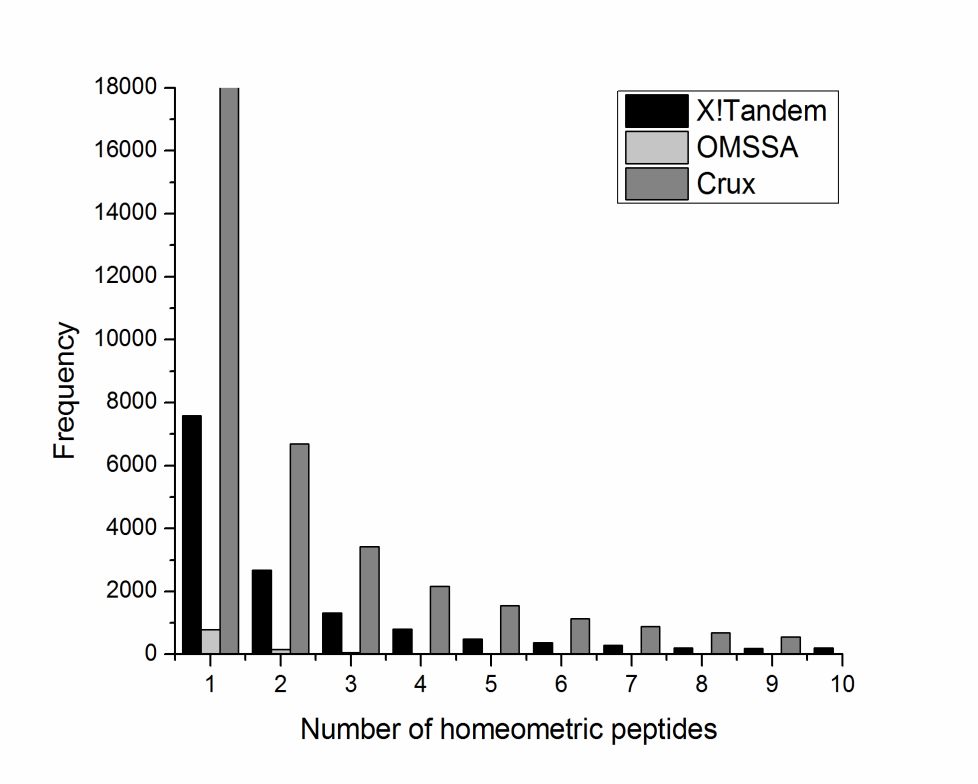

Figure 2.2. Frequency (number) of spectra with 1 to 10 homeometric matches for K106 k-permuted decoy databases across the three database search programs (X! Tandem, OMSSA, and Crux).

2.8 TABLES

Table 2.1. Peptide detection significance levels using ideal simulated spectra of the 103 peptides with and without any post-translational modifications (PTMs) and all *b*- and *y*-ions including neutral mass losses against a standard target database across database search programs (OMSSA, X! Tandem, and Crux).

Program	**PTMs**	**Significance[a]**							**$P \leq 10^{-2}$[b]**
		0	**1**	**2**	**3**	**4**	**5**	**≥6**	
X! Tandem	None	0	0	4	4	2	6	58	74
	Amidation	0	0	0	0	0	0	9	9
	Oxidation	0	0	0	0	0	0	1	1
	Pyroglutamination	0	0	1	0	1	1	1	4
	Phosphorylation	0	0	0	0	0	1	3	4
	N-terminal acetylation	0	0	0	0	0	0	0	0
OMSSA	None	0	0	0	0	0	0	79	79
	Amidation	0	0	0	0	0	0	9	9
	Oxidation	0	0	0	0	0	0	1	1
	Pyroglutamination	0	0	1	2	1	0	0	4
	Phosphorylation	0	0	0	0	0	0	4	4
	N-terminal acetylation	0	0	0	0	0	0	5	5
Crux	None	2	5	12	52	3	1	2	70
	Amidation	0	0	2	3	2	0	2	9
	Oxidation	0	0	0	1	0	0	0	1
	Pyroglutamination	0	1	1	0	1	0	1	3
	Phosphorylation	0	0	1	2	0	0	1	4
	N-terminal acetylation	0	0	0	3	1	0	1	5

[a]Significance threshold (t) for matched to be considered significant at *E*- or *p-value* < 1×10^{-t} (t = 0 to >= 6). [b]Cumulative number of peptides with *E*- or *p-value* < 1×10^{-2}.

Table 2.2. Peptide detection significance levels using experimental spectra of the 103 peptides with and without any post-translational modifications (PTMs) against a standard target database across database search programs (OMSSA, X! Tandem, and Crux).

Program	**PTMs**			**Significance[a]**							**Cum N[b]**
		Miss[c]	**Inc[d]**	**0**	**1**	**2**	**3**	**4**	**5**	**≥6**	**$P \leq 10^{-2}$**
X! Tandem	None	0	0	1	8	11	15	16	11	18	71
	Amidation	0	0	0	0	0	0	0	0	9	9
	Oxidation	0	0	0	0	0	0	0	0	1	1
	Pyroglutamination	0	0	0	0	1	0	1	1	1	4
	Phosphorylation	0	0	0	0	0	0	0	1	3	4
	N-terminal acetylation	0	5	0	0	0	0	0	0	0	0
OMSSA	None	0	0	0	0	1	2	1	3	73	80
	Amidation	1	0	1	0	0	0	1	0	6	7
	Oxidation	0	0	0	0	0	0	0	0	1	1
	Pyroglutamination	0	0	0	0	0	0	1	0	3	4
	Phosphorylation	0	0	0	0	0	0	0	0	4	4
	N-terminal acetylation	0	1	0	0	0	0	0	0	4	4
Crux	None	0	0	9	8	9	44	1	0	9	63
	Amidation	0	0	0	1	5	1	1	0	1	8
	Oxidation	0	0	0	0	0	1	0	0	0	1
	Pyroglutamination	0	0	1	1	0	2	0	0	0	2
	Phosphorylation	0	0	0	2	2	0	0	0	0	2
	N-terminal acetylation	0	0	0	1	1	2	0	1	0	4

[a]Significance threshold (t) for matched to be considered significant at *E*- or *p-value* < 1 x 10^{-t} (t = 0 to >= 6). [b]Cumulative number of peptides with *E*- or *p-value* < 1 x 10^{-2}. [c]Number of peptides missed by program. [d]Number of peptides with incorrect post-translational modification assignment.

Table 2.3. Performance of the target and alternative k-permuted decoy databases used with the X! Tandem database search program using spectra from 80 unmodified neuropeptides.

Database[a]	**Indicator**	**Significance Levels of the Permutation *p-values*[b]**							**Cum. Num. of Peptides[c]**	
		0	**1**	**2**	**3**	**4**	**5**	**≥6**	**≥10^{-2}**	**≥10^{-4}**
Target	*E-value*	1	8	11	15	16	12	17	71	45
$K10^3$	# ions	0	0	76	4	0	0	0	80	0
	Hyperscore	0	0	76	4	0	0	0	80	0
	Convolution	0	8	70	2	0	0	0	72	0
	E-value	0	0	76	4	0	0	0	80	0
$K10^4$	# ions	0	0	0	80	0	0	0	80	0
	Hyperscore	0	0	0	80	0	0	0	80	0
	Convolution	0	5	44	31	0	0	0	75	0
	E-value	0	0	0	80	0	0	0	80	0
$K10^5$	# ions	0	0	0	0	80	0	0	80	80
	Hyperscore	0	0	0	0	80	0	0	80	80
	Convolution	0	3	36	32	9	0	0	77	9
	E-value	0	0	0	0	80	0	0	80	80
$K10^6$	# ions	0	0	0	0	1	79	0	80	80
	Hyperscore	0	0	0	0	0	80	0	80	80
	Convolution	0	4	30	36	5	5	0	76	10
	E-value	0	0	0	0	0	80	0	80	80

[a]Target: database of 236 neuropeptide sequences; $K10^3$: k-permuted decoy database size of 236,000 peptides; $K10^4$: k-permuted decoy database size = 2,360,000 peptides; $K10^5$: k-permuted decoy database size = 23,600,000 peptides; $K10^6$: k-permuted decoy database size = 236,000,000 peptides. [b]Significance threshold (t) for target spectrum to be considered significant at significance thresholds $< 1 \times 10^{-t}$ (t = 0 to >= 6). [c]The cumulative number of peptides at 1×10^{-2} and 1×10^{-4} thresholds.

Table 2.4. Performance of the target and alternative k-permuted decoy databases used with the Crux database search program using spectra from 80 unmodified neuropeptides.

Database[a]	**Indicator**[b]	**Significance Levels of the Permutation *p-values*[c]**							**Cum. Num. of peptides**[d]	
		0	**1**	**2**	**3**	**4**	**5**	**≥6**	**≥10^{-2}**	**≥10^{-4}**
Target	*p-value*	9	8	9	44	1	0	9	63	10
$K10^3$	# ions	0	2	78	0	0	0	0	78	0
	XCorr	3	11	66	0	0	0	0	66	0
	Sp	0	2	78	0	0	0	0	78	0
	ΔCn	3	11	66	0	0	0	0	66	0
$K10^4$	# ions	0	0	1	79	0	0	0	80	0
	XCorr	3	10	14	53	0	0	0	67	0
	Sp	0	0	1	79	0	0	0	80	0
	ΔCn	3	10	14	53	0	0	0	67	0
$K10^5$	# ions	0	0	0	1	79	0	0	80	79
	XCorr	3	10	8	23	36	0	0	67	36
	Sp	0	0	0	1	79	0	0	80	79
	ΔCn	3	10	8	23	36	0	0	67	36
$K10^6$	# ions	0	0	0	0	2	78	0	80	80
	XCorr	3	10	9	19	22	17	0	67	39
	Sp	0	0	0	0	4	76	0	80	80
	ΔCn	3	10	9	19	22	17	0	67	39

[a]Target: database of 236 neuropeptide sequences; $K10^3$: k-permuted decoy database size of 236,000 peptides; $K10^4$: k-permuted decoy database size = 2,360,000 peptides; $K10^5$: k-permuted decoy database size = 23,600,000 peptides; $K10^6$: k-permuted decoy database size = 236,000,000 peptides. [b]# ions: permutation *p-values* computed for the number of matched *b*- and *y*-ions. XCorr: permutation *p-values* computed from the XCorr scores of the matches. Sp: permutation *p-values* computed from the Sp scores of the matches. ΔCn: permutation *p-values* computed using X! Tandem ΔCn. [c]Significance threshold (t) for matched to be considered significant at *p-value* < 1 x 10^{-t}. [d]Cumulative number of peptides with *p-values* thresholds of 1 x 10^{-2} and 1 x 10^{-4}.

Table 2.5. Performance of the target alternative k-permuted decoy databases used with the OMSSA database search program using spectra from 80 unmodified neuropeptides.

Database[a]	**Indicator**[b]	**Significance Levels of the Permutation *p-values*[c]**							**Cum. Num. of Peptides**[d]	
		0	**1**	**2**	**3**	**4**	**5**	**≥6**	**≥10^{-2}**	**≥10^{-4}**
Target	*E-value*	0	0	1	2	1	3	73	80	77
$K10^3$	# ions	0	2	78	0	0	0	0	78	0
	Lambda	0	9	71	0	0	0	0	71	0
	p-value	0	2	78	0	0	0	0	78	0
	E-value	0	2	78	0	0	0	0	78	0
$K10^4$	# ions	0	0	1	79	0	0	0	80	0
	Lambda	0	5	11	64	0	0	0	75	0
	p-value	0	0	1	79	0	0	0	80	0
	E-value	0	0	1	79	0	0	0	80	0
$K10^5$	# ions	0	0	0	0	80	0	0	80	80
	Lambda	0	5	8	24	43	0	0	75	43
	p-value	0	0	0	0	80	0	0	80	80
	E-value	0	0	0	0	80	0	0	80	80
$K10^6$	# ions	0	0	0	0	2	78	0	80	80
	Lambda	0	5	8	17	18	32	0	75	50
	p-value	0	0	0	0	0	80	0	80	80
	E-value	0	0	0	0	0	80	0	80	80

[a]Target: database of 236 neuropeptide sequences; $K10^3$: k-permuted decoy database size of 236,000 peptides; $K10^4$: k-permuted decoy database size = 2,360,000 peptides; $K10^5$: k-permuted decoy database size = 23,600,000 peptides; $K10^6$: k-permuted decoy database size = 236,000,000 peptides. [b]# ions: permutation *p-values* computed for the number of matched *b*- and *y*-ions. Lambda: permutation *p-values* computed from the Poisson mean of matches. *p-value*: permutation *p-values* computed from the *p-value* reported by the OMSSA for the matches. *E-value*: permutation *p-values* computed using OMSSA *E-values*. [c]Significance threshold (t) for matched to be considered significant at *p-value* < 1 x 10^{-t}. [d]Incorrect: the program provided an incorrect match. [e]Cumulative number of peptides with *p-value* < 1 x 10^{-2}. [f]Cumulative number of peptides with *p-value* < 1 x 10^{-4}.

Table 2.6. Computation times given in seconds for search of 80 unmodified spectra against different databases using a single process Intel® Core™ i7-3770 CPU @ 3.40GHz.

Database[a]	**Database Search Program**		
	Crux	**OMSSA**	**X! Tandem**
Target	5	11	1
$K10^3$	7	56	41
$K10^4$	61	915	476
$K10^5$	200	1220	467
$K10^6$	2162	24475	5196

[a]Target: database of 236 neuropeptide sequences; $K10^3$: k-permuted decoy database size of 236,000 peptides; $K10^4$: k-permuted decoy database size = 2,360,000 peptides; $K10^5$: k-permuted decoy database size = 23,600,000 peptides; $K10^6$: k-permuted decoy database size = 236,000,000 peptides.

CHAPTER III: IDENTIFICATION OF BEST INDICATORS OF PEPTIDE-SPECTRUM MATCH USING A PERMUTATION RESAMPLING APPROACH

3.1 NOTES AND ACKNOWLEDGMENTS

The material presented in this chapter is reproduced with permission (http://www.worldscientific.com/page/authors/author-rights) from the preprint version of an article "Akhtar, M. N.; Southey, B. R.; Andren, P. E.; Sweedler, J. V.; Rodriguez-Zas, S. L. Identification of best indicators of peptide-spectrum match using a permutation resampling approach" published in Journal of Bioinformatics and Computational Biology (2014). This work was completed in Dr. Sandra Rodriguez-Zas Bioinformatics Laboratory at University of Illinois Urbana-Champaign (USA) in collaboration with Dr. Sweedler at University of Illinois Urbana-Champaign (USA) and Dr. Andren at University of Uppsala (Sweden). The work is focused on identification of match indicators providing better performance within and across programs using significance values estimated for the neuropeptide identifications from the database search programs using terminal permutations of peptides in the target database. The support of NIH (Grant Numbers: R21 DA027548, P30 DA018310 and R21 MH096030) is greatly appreciated.

3.2 ABSTRACT

Various indicators of observed-theoretical spectrum matches were compared and the resulting statistical significance was characterized using permutation resampling. Novel decoy databases built by resampling the terminal positions of peptide sequences were evaluated to identify the conditions for accurate computation of peptide match significance levels. The methodology was tested on real and manually curated tandem mass spectra from peptides across a wide range of sizes. Indicators from complementary database search programs were profiled. The permuted decoy databases improved the calculation of the peptide match significance compared to the approaches currently implemented in the database search programs that rely on distributional assumptions. Permutation tests using *p-values* obtained from software-dependent matching scores and *E-values* outperformed permutation tests using all other indicators. The higher overlap in matches between the database search programs when using end permutation compared to existing approaches confirm the superiority of the end permutation method to identify peptides. The combination of effective match indicators and the end permutation method is recommended for accurate detection of peptides.

3.3 INTRODUCTION

Mass spectrometry discovery has revolutionized proteomic research enabling the characterization and quantification of hundredths of peptides from samples ranging in size and complexity.[74-79] In tandem mass spectrometry (MS/MS) experiments, the peptides present in the sample can be identified by sequence database search programs.[16, 88] These programs attempt to match the fragment ions from the observed spectra with the fragment ions from theoretical spectra generated from the known or predicted peptide sequences in the target database. Each observed-theoretical spectra match is assigned scores that reflects the similarity between both spectra. Subsequently, these scores are converted into a measure of the statistical evidence supporting the match.[41, 46]

Two related components, the match score and the statistical significance assigned to the score, influence the capability to detect peptides. Database search software differ in the algorithms and assumptions to assess the observed-theoretical spectra match leading to different mating score indicators (e.g., number of matched fragment ions, cross-correlation) and different methods to assess statistical significance of the match. The comparative effectiveness of the scores to capture the match has not been evaluated.

One commonly used approach to convert specific observed-theoretical spectra match score into a statistical significance value encompasses fitting a specific parametric distribution to all the match scores attained from the target database[43, 45] or from decoy peptides generated from the target database matches.[56] Alternatively, significance values can be obtained in a non-parametric fashion from the decoy peptides.[69] A previous comparative study of the database search programs

demonstrated that, for some peptides, detection using significance value estimation approaches implemented in the database search programs remains challenging.[16] This situation can be traced back to the low significance levels obtained with existing approaches particularly for short peptides under 15 amino acids in length.[16]

The challenges of peptide identification using existing approaches include false negatives due to match significance levels that do not surpass the minimum detection threshold, false positives due to incorrectly spectra match surpassing the minimum threshold, and missed peptides due to sample complexity leading to multiple peptides present in the single tandem spectrum (also known as chimeric spectra).[16] The bias introduced by existing approach has major impact in small peptides. These peptides are unlikely to be identified at high significance levels by most database search programs due limited number of fragment ions to accumulate high matching scores.[16, 20, 58] Also, tandem spectra that have incomplete fragmentation and noise peaks can result in matches with low scores that can be indifferent to the random matches, thus, resulting in low significance levels.[16, 80] Likewise, increases in the effective search database size (such as those rising from the consideration of post-translational modifications) can reduce the sensitivity of the algorithms to detect peptides at accurate significance levels.[20]

In the target-decoy approach, observed spectra are matched to theoretical spectra from reverted or reshuffled sequences from the target database together with the original target sequences.[65] The target-decoy approach aims at avoiding stringent significant threshold to control for multiple testing across peptides.[72, 73] However, for small peptides, most decoy database construction methods produce few spectra that have more extreme matches that artificially inflates the significance levels. Other decoy databases construction methods that exploit the capability of

resampling approaches to generate null hypothesis while controlling the experiment-wise error rate should be evaluated.

The aims of this study were: (1) to compare indicators of observed-theoretical spectra matches and characterize the accuracy of the resulting statistical significance using permutation testing, (2) to develop novel decoy databases including resampling of terminal positions in the peptide sequence and identify the conditions for accurate computation of match significance levels, and (3) to demonstrate the application of the novel decoy approach using popular database search programs.

3.4 THEORETICAL-OBSERVED SPECTRA MATCH INDICATORS

Table 3.1 lists the observed-theoretical spectrum match indicators evaluated and corresponding database search programs: Crux (version 1.37),[44] OMSSA (version 2.1.8),[43] and X! Tandem (version 2013.02.01.1).[45] These programs were selected because their open source nature allowed the retrieval of intermediate match indicators through modification of the source code.

Database search specifications were: (1) mass type: monoisotopic; (2) fragment ion charge: default values; "mz-bin-width": 0.3 (Crux); (3) no post-translational modifications; (4) enzyme: "whole protein" (OMSSA) or custom cleavage site to avoid cleavage of the provided neuropeptide database (Crux and X! Tandem); (5) precursor ion tolerance: 1.5 Da; (6) fragment ion tolerance: 0.3 Da (OMSSA and X! Tandem); and (7) OMSSA "ht": 8 to consider only those database peptides that had one or more fragment ion matching including one of top 8 highest fragment ion peaks in the observed spectrum. The selected specifications follow program settings previously used to evaluate the ability of the database search programs to identify peptides.[16]

3.5 OBSERVED SPECTRA, TARGET AND DECOY DATABASES

The performance of alternative indicators to assign the statistical significance to spectra matches was investigated on a murine linear ion trap (LTQ) tandem spectra dataset.[25] Spectra and peptide identification were obtained from the SwePep database (http://www.swepep.org).[25] The tandem spectra dataset consisted of 80 observed tandem spectra from neuropeptides without post-translational modifications. The majority of the peptides (92%) had precursor charge states +2 or +3. The target database included the 80 peptides with observed spectra studied and all other peptides that could have been produced from the known 95 mouse prohormones including those that produced the 80 peptides studied. The exhaustive list of target peptides was obtained from the PepShop[32] database (http://stagbeetle.animal.uiuc.edu/pepshop) including information from the SwePep, UniProt,[24] and NeuroPred.[85]

To understand the performance of the software under best conditions, optimal spectra were simulated for the peptides in the target database using corresponding precursor charge states. For each spectrum, all *b*- and *y*- fragment ions with +1 charge state were simulated with uniform intensity. Additional peaks due to loss of single ammonia or water molecule were simulated when the *b*- or *y*-ions sequence contained a water or ammonia loosing amino acids anywhere in the sequence.[16]

The characterization of the significance of the spectral match based on various indicators relied on a decoy database generated using permutation.[84] A single target database that was used for all database search programs was created by selecting all peptides within 12 Da (corresponding to 3 m/z ion tolerance with a +4 charge state) of the precursor mass for each tandem spectrum. This

mass limit results from the database search programs preselecting candidate peptides based on peptide mass and user-defined mass tolerances. Permutations of each target candidate sequence residues at the N- and C-terminal ends were used to populate the decoy database. This terminal permutation generated decoy peptides that were more similar to their target peptides yet disrupted the pattern of *b*- and *y*-fragment ions that are used in matching the observed and theoretical spectra. The terminal regions were selected because it was considered that the ions from the terminal regions had better sensitivity than the ions from the central region of peptide. Leucine and isoleucine were treated as the same amino acid in all permutations and comparisons between candidate and permuted sequence.

From the termini permutation strategy, three decoy databases: Ends1, Ends2 and Ends3 were evaluated. Ends1 encompasses 236*(19 N-terminal amino acids)*(19 C-terminal amino acids) = 236*360 = 84,960 decoy peptides; Ends2 encompasses 236*(19*19 N-terminal amino acids)*(19*19 C-terminal amino acids) = 236*130320 = 30,755,520 decoy peptides; and Ends3 encompasses 236*(19*19*19 N- terminal amino acids)*(19*19*19 C-terminal amino acids) = 236*47,045,880 =1,120,027,680 decoy peptides. Separate permuted databases were created for each observed spectra in Ends3 due to inability of the database search programs to adequately handle the size of the permuted decoy database. The target database was appended to each of the Ends decoy databases for the combined target-decoy search strategy. The merging of the target and decoy databases provided unbiased *p-value* estimates and avoided zero *p-value*s.[84]

For each observed-theoretical spectra match indicator, the permutation *p-values* were computed as the relative frequency of the sum of the matches in the target-decoy database that had indicator values equal or better than the observed-target spectra matches. A Bonferroni adjusted

threshold *p-value* < 1 x 10^{-4} based on a 1% experiment-wise error rate (0.01/80 ≈ 1 x 10^{-4}) was used to compare performance of the different indicators. A sensitivity analysis enabled the assessment of the impact of the *p-value* threshold on the capability of match indicators to detect the peptides. The limited number of observed and annotated spectra prevented unbiased analysis using receiver operating characteristic (ROC) curve.

3.6 RESULTS AND DISCUSSION

A threefold-strategy was used to characterize the performance of spectra match indicators from database search programs to detect peptides. First, optimal simulated spectra were searched against the target database to obtain a baseline performance in the absence of data quality issues such as presence of noise peaks, missing signal peaks, and low signal-to-noise ratio. Second, real spectra were searched against the target database to study the influence of data quality issues on peptide detection significance levels relative to the baseline performance. Third, the performance of the match indicators to detect peptides in realistic scenarios using End-permuted decoy databases was demonstrated.

PEPTIDE DETECTION BENCHMARKS USING OPTIMAL AND REAL SPECTRA AGAINST THE TARGET DATABASE

Table 3.2 summarizes the number of peptides detected by the three database search programs at various significance *E*- or *p-value* thresholds when optimal uniform simulated spectra and real tandem mass spectra were searched against the target database.

For the ideal simulated spectra, the three programs accurately detected all peptides at *E*- or *p-value* < 1 x 10^{-1}. At *E*- or *p-value* < 1 x 10^{-4}, the Crux, OMSSA, and X! Tandem detected 9 (11.25%), 80 (100%), and 72 (90.0%) peptides, respectively. The significance levels of the X! Tandem *E-values* increased linearly with increase in peptide length and only peptides greater than 8 amino acids in length (hyperscore > 40) reached a significance level of *E-value* < 1 x 10^{-4}. OMSSA *E-values* were less correlated with peptide length or number of matched *b*- and *y*-ions. The minimum *E-value* was 1 x 10^{-6} and corresponded to an 11 amino acid-long peptide that had a +2 precursor charge state spectrum. The lower significance level of Crux peptide matches, relative to the OMSSA and X! Tandem, have been confirmed previously.[16] At a less stringent threshold *p-value* < 1 x 10^{-2}, Crux identified 73 (91.25%) peptides with seven peptides between 7 to 14 amino acids in length undetected.

Crux, OMSSA, and X! Tandem correctly matched 10 (12.5%), 77 (96.35), and 45 (56.3%) real spectra, respectively, at *E*- or *p-value* < 1 x 10^{-4}. A large number of peptides (44) were detected with a *p-value* < 10^{-3} indicating the previously noted difficulty of obtaining significant matches with Crux.[16] The spectra quality features such as missing peaks, noise peaks and low intensity peaks tended to reduce the positive correlation that was observed between peptide length and *E-value* in the optimal simulated scenario.

Higher number of Weibull points (XCorr scores) were correlated with more significant *p-values* in Crux.[16] Consistent with prior work, the increase in the number of Weibull points from 10^3 to 10^4, and 10^5 resulted in 24 and 10 more peptides that reached *p-value* < 1 x 10^{-4} relative to the 10^3 scenario, respectively. However, 17 and 40 more peptides had *p-value* > 1 x 10^{-2} with 10^4 and 10^5 Weibull points, respectively, than with 10^3 points (data not shown). Further investigation

uncovered that peptides that did not reach the significance threshold were affected by the "mz-bin-width" (fragment ion tolerance) parameter. Increasing the "mz-bin-width" values from 0.3 to 1.0005 increased XCorr scores, and consequently, reduced the number of peptides that had *p-value* $> 1 \times 10^{-2}$ (**figure 3.1**). Thus, the 0.3 specification appears to provide more conservative results. However, to use comparable search specification for the three database search programs, from this point onwards, all Crux results were calculated using the more conservative 0.3 "mz-bin-width".

PEPTIDE DETECTION USING REAL SPECTRA AGAINST THE END DECOY DATABASE

The detection of peptides from observed real spectra when matched against the End-permuted decoy database improved relative to the standard comparison against a target database. **Figure 3.2** depicts the distribution of the effective database size corresponding to each observed spectra for the three database search programs when two (Ends2) or three (Ends3) terminal residues were permuted. The patterns in these box plots showed that X! Tandem evaluated more decoy sequences than the Crux and OMSSA.

For each peptide, some matches of the observed spectrum against the decoy database spectra were indistinguishable from each other in terms of all indicators (e.g., the number of matched fragment ions, XCorr score, and Sp score). This is because for each peptide, the Ends2 and Ends3 decoy databases had dimer and trimer residue combinations with similar total monoisotopic masses. These numerically indistinguishable matches were counted as one when calculating the permutation *p-values* to avoid biases towards any one database search program.

Table 3.3 summarizes the number of peptides matched at different $\log_{10}$-transformed permuted *p-value* significant levels across match indicators and database search programs for the Ends1, Ends2, and Ends3 decoy databases.

X! Tandem

The level of significance of the matches to the decoy databases increased from Ends1 to Ends2 and stabilized between Ends2 and Ends3 decoy databases (**Table 3.3**). The Ends2 and Ends3 decoy databases detected 34.95 to 38.70% more peptides than the target database. Overall, the X! Tandem indicator convolution score had the lowest detection rate among all indicators suggesting that the convolution score alone is inadequate to discriminate between true target and false decoy matches. Detections and significance levels were similar for the hyperscore and *E-value* indicators. Furthermore, detection rate was comparable between hyperscore and the number of matched ions across the three End decoy databases. End decoy databases improved peptides detection relative to the target database for number of matched ions, hyperscore and *E-value* indicators.

The peptides that were not detected by the hyperscore were also not detected by the number of matched ion indicator. The decoy database size was not correlated with the significance level or capability to detect the peptide. Of the undetected peptides, 2 peptides were not detected by the Ends2 and Ends3 databases. Meanwhile five undetected peptides in the Ends2 database were significant with the Ends3 database, four other peptides that were significant in the Ends2 database were not detected (became non-significant) in the Ends3 decoy database. The non-significant peptides in the Ends3 database were either non-significant or marginally significant in the target database.

Table 3.4 summarizes the number of peptides detected in the target and Ends3 decoy databases, target only, Ends3 only, and missed by both databases when the number of matched ions and hyperscore indicators are considered. The Ends3 decoy database detected most peptides (42 out of 45) that were significant in the target database in addition to the 32 peptides that were missed by the standard target database. The performance of the number of matched ions and hyperscore was comparable. The higher significance of the matches resulting from the consideration of the hyperscore relative to all other X! Tandem indicators can be attributed to the use of peak intensity in the scoring and the theoretical spectrum synthesis process.[20]

Crux

Peptide detection and significance levels were similar for the XCorr and ΔCn across Ends2 and Ends3 decoy databases. The XCorr and ΔCn detected 33 (41.25%) and 35 (43.75%) peptides in the Ends2 and Ends3 decoy databases, respectively (**Table 3.3**). The lower peptide detection rate of XCorr and ΔCn with decoy databases indicates that XCorr and ΔCn are less suitable than the other indicators (Sp and number of ions). Overall, the Sp indicator identified 2 and 4 more peptides (*p-value* < 1 x 10^{-4}) than the number of matched ions indicator in Ends2 and Ends3, respectively (**Table 3.3**).

Combining the number of matched ions or Sp indicators with the End decoy databases improved the peptide detection relative to the target database alone. The Ends2 and Ends3 databases had 67.5 to 83.75% peptide detection rate compared to 12.50% with the target database with both indicators. The number of matched ion indicator missed more peptides (23) than the Sp indicator (19). The Ends3 permuted database detected 51 peptides missed by the standard target database using Sp indicator (**Table 3.4**).

OMSSA

Table 3.3 summarizes the $\log_{10}$-transformed *p-values* for the OMSSA match indicators: number of matched ions, lambda, Poisson *p-value*, and *E-value*. Detections and significance levels were identical for the Poisson *p-value* and *E-value* indicators, and only *E-value* indicator would be considered in further discussion. The lambda indicator overall detected lower number of peptides than the number of matched ion and *E-value* indicators suggesting that the lambda alone is inadequate to discriminate between target and decoy matches. The Ends2 and Ends3 decoy databases provided further discrimination between the number of matched ions and *E-value* indicators, with significance levels and peptide detection rate in the decoy database higher than the target database when the *E-value* indicators was considered. The *E-value* indicator provided more true detections across significance thresholds than the number of matched ions and lambda indicators.

Comparison among Database Search Indicators

Table 3.4 lists the number of peptides identified by the target and Ends3 decoy, target only, Ends3 decoy only, and not identified by either database when the number of matched ions and *E-value* indicators are considered. Meanwhile the number of ions and *E-value* indicators detected 3 peptides using the Ends3 decoy database that were missed by the target database, these indicators detected 10 and 7 peptides, respectively using the target database that were missed by the decoy database. Approximately, 88% peptide detections were shared by the target and Ends3 databases using the *E-value* indicator.

3.7 COMPARISON OF SPECTRA MATCH INDICATORS AND DATABASE SEARCH SOFTWARE

Figure 3.3 depicts the number of peptides detected by one, two or all three database search programs when the number of matched ion and best score indicator from each of the three programs was used to compute the *p-value.* The best score indicator was defined as the indicator that exhibited the highest difference between the target and decoy peptides. The best spectra match indicators were *E-value* for OMSSA, hyperscore for X! Tandem, and Sp for Crux.

The Ends decoy databases supported higher consensus among the three programs when compared to the target database. For the Ends3 decoy database, all three programs detected slightly less peptides together when considering the number of matched ions compared to the best indicator (50 vs. 56). A similar number of peptides were detected by any two programs using the number of matched ions than the best score indicator (72 vs. 73). OMSSA and Crux detected more peptides with best indicator than the number of matched ion indicator and X! Tandem detected similar peptides with the number of matched ions and the hyperscore. Using either the number of matched ions or best score indicator, X! Tandem detected more peptides than OMSSA and Crux and OMSSA detected more peptides than Crux.

The computational time of the searches was calculated on a computer with 3.40 GHz Intel Core i7-3770 processor. Searching the target database only using Crux (using 1000 Weibull points), X! Tandem and OMSSA averaged 1.14, 0.013, and 0.14 seconds per spectrum, respectively. Crux averaged 0.04, 3.54 and 40.65 seconds for Ends1, Ends2 and Ends3 decoy databases, respectively. X! Tandem averaged 0.15, 6.54, and 116.26 seconds per spectrum for

Ends1, Ends2 and Ends3 decoy databases, respectively. OMSSA averaged 0.34, 21.72, and 604.00 seconds per spectrum for Ends1, Ends2 and Ends3 decoy databases, respectively. The longer search time for the X! Tandem and OMSSA using the Ends3 decoy database relative to Ends2 database could be due to the searching of separate decoy databases for each spectrum in addition to the larger database size of the Ends3 decoy database. Furthermore, the comparisons of the peptide detection rate between the Ends2 and Ends3 database suggest that detection performance similar to the Ends3 database could be obtained using a smaller random sample of the decoys in the Ends3 database. Overall, the dramatic improvement in the peptide identification highlights the efficacy of the terminal residue permutation decoy database.

3.8 CONCLUSIONS

The present study demonstrated that the spectra match indicators Sp (Crux), hyperscore (X! Tandem) and *E-value* (OMSSA) with a terminal residue permutation decoy database enabled effective detection of peptides compared to target database. The Ends decoy databases improved the consensus among database search programs to identify peptides. The End decoy databases can be integrated to other database search programs. The new candidate decoy peptides resulting from the permutation can also be used to discover novel peptides.

In the present study, Ends decoy databases were generated from subset of target database peptides that were within 12 Da of the observed spectra precursor masses since database search programs initially filter candidate peptides based on precursor mass. The approach can be extended to any number of peptides, types of peptides and other database search programs. This could be accomplished by generating the required number of permuted peptides from peptide-spectrum

matches (PSMs) obtained by searching observed spectra against the target database using the desired database search program.

3.9 FIGURES

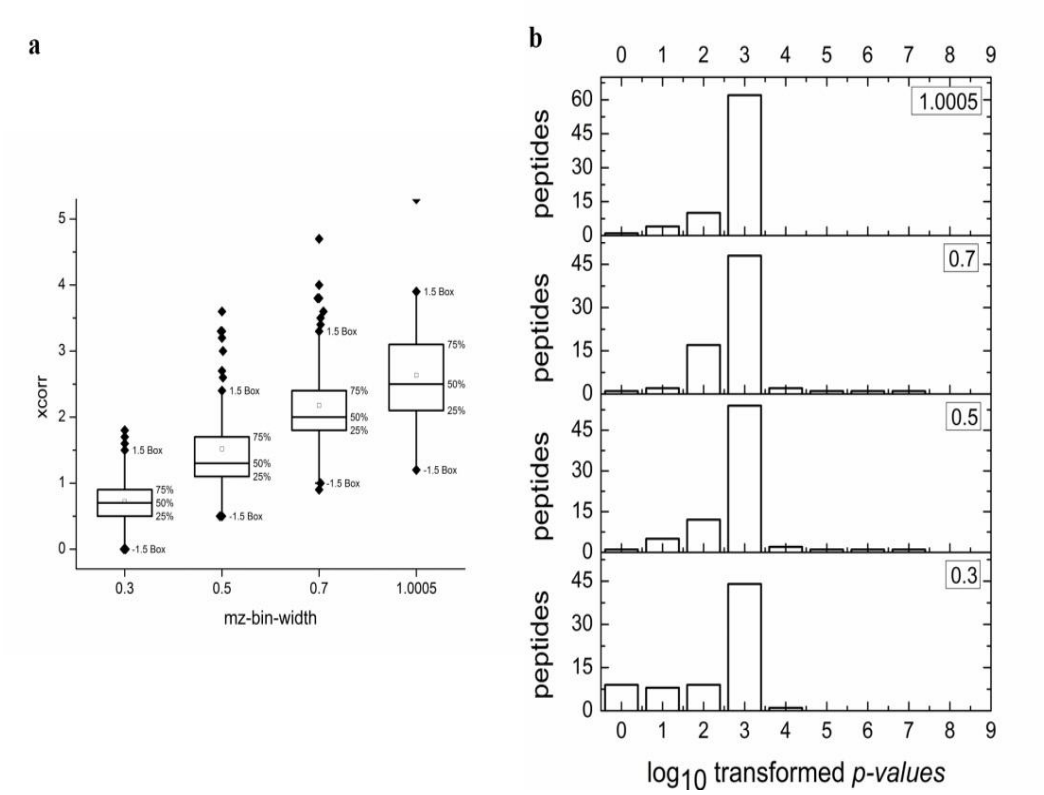

Figure 3.1. Box plots of Crux XCorr scores (a) and number of peptides correctly identified at different -1*$\log_{10}$-transformed Weibull *p-values* (b) using "mz-bin-width" values of 0.3, 0.5, 0.7, and 1.0005.

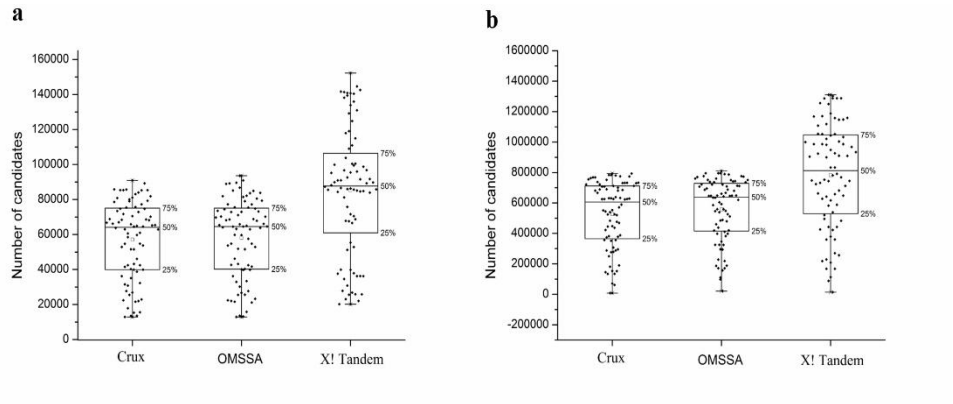

Figure 3.2. Box plots depicting the distribution of number of candidate decoy peptides within precursor mass tolerance per queried observed peptide considered by Crux, OMSSA, and X! Tandem for the (a) Ends2 and (b) Ends3 permuted decoy databases.

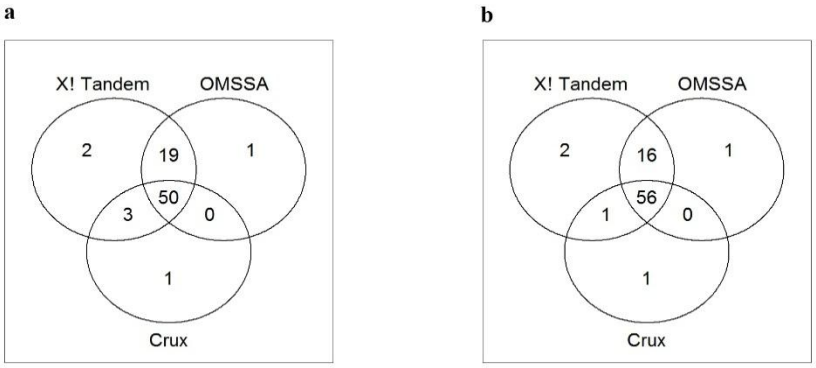

Figure 3.3. Distinct and shared number of peptide detected in the Ends3 decoy database using a) the number of matched ions or b) the best indicator for each database search program (OMSSA *E-value*, Crux Sp score, and X! Tandem hyperscore).

3.10 TABLES

Table 3.1. Crux, X! Tandem, and OMSSA match indicators used.

Programs	**Indicators**
Crux	Number of matched *b*- and *y*-fragment ions SEQUEST preliminary (Sp) score Cross-correlation (XCorr) score DeltaCn (ΔCn) score *p-value*: computed from the Weibull distribution using 10^3 XCorr scores
X! Tandem	Number of matched *b*- and *y*-fragment ions Convolution score Hyperscore *E-value:* computed assuming hypergeometric distribution for hyperscores
OMSSA	Number of matched *b*- and *y*-fragment ions Lambda or Poisson mean Poisson *p-value* *E-value*: Poisson *p-value* multiplied by effective database size

Table 3.2. Number of peptides matched at various significance levels of the log_{10}-transformed *E*- or *p-values* when the optimal simulated spectra and real tandem spectra were searched against the standard target database.

Program	**Spectra**	**Log_{10}-transformed *p-values***							**Peptides (%) at < 1 x 10^{-4}**
		0[a]	**1**	**2**	**3**	**4**	**5**	**≥6**	
Crux	Optimal	2	5	12	52	3	1	5	11.3
	Real	9	8	9	44	1	0	9	12.5
OMSSA	Optimal	0	0	0	0	0	0	80	100.0
	Real	0	0	1	2	1	3	73	96.3
X! Tandem	Optimal	0	0	4	4	2	6	64	90.0
	Real	1	8	11	15	16	11	18	56.3

[a]Significance threshold (t) for matches to be significant at *p-value* <1 x 10^{-t}.

Table 3.3. Number of peptides detected by spectra match indicators from database search programs across log_{10}-transformed *p-values* levels of the computed using the End decoy databases.

Programs	**Database**[a]	**Indicators**	**Log_{10}-transformed *p-values***							**Pep.**
			0[b]	**1**	**2**	**3**	**4**	**5**	**≥6**	**< 1 x 10^{-4}**[c]
X! Tandem	Ends1	# of ions	0	8	72	0	0	0	0	0
		Convolution	0	25	55	0	0	0	0	0
		Hyper/*E-value*	0	9	71	0	0	0	0	0
	Ends2	# of ions	0	0	0	7	65	8	0	73
		Convolution	0	2	20	41	17	0	0	17
		Hyper/*E-value*	0	0	0	4	67	9	0	76
	Ends3	# of ions	0	0	0	6	29	44	1	74
		Convolution	0	0	1	26	31	22	0	53
		Hyper/*E-value*	0	0	0	5	20	51	4	75
Crux	Ends1	# of ions	0	20	60	0	0	0	0	0
		Sp	0	19	61	0	0	0	0	0
		XCorr/ΔCn	4	30	46	0	0	0	0	0
	Ends2	# of ions	0	0	0	15	65	0	0	65
		Sp	0	0	0	13	67	0	0	67
		XCorr/ΔCn	1	6	12	28	33	0	0	33
	Ends3	# of ions	0	1	1	24	27	27	0	54
		Sp	0	1	1	20	28	30	0	58
		XCorr/ΔCn	0	3	17	25	23	12	0	35
OMSSA	Ends1	# of ions	0	16	64	0	0	0	0	0
		Lambda	2	29	49	0	0	0	0	0
		p-value/*E-value*	0	14	66	0	0	0	0	0
	Ends2	# of ions	0	0	0	22	58	0	0	58
		Lambda	0	6	15	25	34	0	0	34
		p-value/*E-value*	0	0	0	11	69	0	0	69
	Ends3	# of ions	0	0	0	10	51	19	0	70
		Lambda	0	0	0	17	43	20	0	63
		p-value/*E-value*	0	0	0	7	33	40	0	73

[a]Ends1: the last one N- and C-terminal amino acids were permuted (decoy peptides: 236*360=84,960); Ends2: the last two N- and C-terminal amino acids were permuted (decoy peptides: 236*130320=30,755,520); Ends3: the last three N- and C-terminal amino acids were permuted (decoy peptides: 47,045,880). [b]Significance threshold (t) for matched to be considered significant at *p-value* < 1 x 10^{-t}. [c]The number of peptides detected at *p-value* < 1 x 10^{-4}.

Table 3.4. Number of peptides detected by spectra match indicators from database search programs using the target and Ends3 decoy databases.

Program	**Indicators**	**Number of peptides detected in Ends3 permuted and target databases**			
		PT[a]	**P**	**T**	**None**
Crux	# of ions	7	47	3	23
	Sp	7	51	3	19
OMSSA	# of ions	67	3	10	0
	E-value	70	3	7	0
X! Tandem	# of ions	42	32	3	3
	Hyperscore	43	32	2	3

[a]PT: peptides detected at *p-value* < 1×10^{-4} in both target and Ends3 databases; P: peptides detected at *p-value* <1×10^{-4} in Ends3 database only; T: peptides detected at *p-value* <1×10^{-4} in the target database only; None: missed peptides (*p-value* > 1×10^{-4}) in both databases.

CHAPTER IV: EVALUATION OF RESAMPLING APPROACH FOR THE TRYPTIC PEPTIDE IDENTIFICATION IN TANDEM MASS SPECTROMETRY EXPERIMENTS USING DATABASE SEARCH APPROACH

4.1 NOTES AND ACKNOWLEDGMENTS

The work presented in this chapter is completed in Dr. Sandra Rodriguez-Zas Bioinformatics Laboratory at University of Illinois Urbana-Champaign (USA) in collaboration with Dr. Sweedler at University of Illinois Urbana-Champaign (USA) and Dr. Andren at University of Uppsala (Sweden). The work is focused on evaluation of the classification performance of the OMSSA *E-value* and k-permuted decoy database using publically available tryptic peptide dataset[20] (http://www.ludwig.edu.au). The support of NIH (Grant Numbers: R21 DA027548, P30 DA018310 and R21 MH096030) is greatly appreciated.

4.2 ABSTRACT

A novel resampling approach was integrated with the OMSSA database search program. Complete peptides sequences that were within a 3 Da precursor mass tolerance of the observed spectrum mass were randomly generated. The approach was tested on 5,806 tryptic tandem mass spectra (http://www.ludwig.edu.au).[20] The performance of the OMSSA's *E-value* indicator and k-permutation decoy database was validated and compared by filtering peptides matches using a 5% false discovery rate estimated from the target database and target-decoy database searches. The conventional receiver operating characteristics (ROC) curve analysis was used to study the tradeoff between true positive rate and false positive rate regardless of any specific threshold. The k-permuted database showed better sensitivity and classification performance relative to the OMSSA *E-value*. A higher peptide detection rate was achieved due to better separation of false negative matches with less number of matched ions and large OMSSA's *E-values* from the true negatives and false positive matches. ROC curves analysis indicated that the k-permuted decoy database had performance comparable with OMSSA *E-value* at various thresholds with an area under the curve (AUC) of 0.95 and 0.94, respectively.

4.3 INTRODUCTION

The bottom-up shotgun approach enables high throughput proteins and peptide identifications from complex protein mixtures using tandem mass spectrometry (MS/MS).[14] In the shotgun approach, proteins in the complex mixture are enzymatically digested into peptides usually with trypsin. The resulting peptides are separated using techniques such as reverse phase chromatography and subsequently introduced into the mass spectrometer. These peptides are ionized and the mass-to-charge (*m/z*) ratios are measured to generate the first MS scan. The peptide ions are further fragmented inside the mass analyzers and the m/z values of the fragments are measured to generate MS/MS (tandem) spectra.[41]

Database search approaches are commonly used to infer the amino acid sequences corresponding to the acquired tandem spectra[41] including SEQUEST,[89] Mascot,[46] OMSSA,[43] X!Tandem,[45] and Crux.[44] For any observed experimental spectra, these programs predict the peptides resulting from the enzymatic digestion of protein sequences compiled in a database, generate *in silico* theoretical spectra from the peptides that fall within mass tolerance of the observed spectra, and match the observed spectra against theoretical spectra. Different database search programs use different scoring schemes to rank the observed-theoretical spectra matches and typically the best match or hit is listed as the most likely identifier peptide.[41] The database search programs can be grouped into probabilistic and empirical based based on the scoring schemes.[67]

The identification of peptides from tandem mass spectra remains challenging.[16, 90] The correct peptide identification (true positive) rate is reported to range from 5 to 50%.[20, 65, 67] The

remaining spectral assignments are either false positive (incorrect), or false negatives (missed).[90] Factors that could lower the true positive rate include presence of low informative product ions due small size of fragmented peptides,[16, 20, 91] incomplete fragmentation,[16, 39, 80] high intensity noise peaks relative to the signal peaks,[26] effective database size,[49] and chimeric events due to the fragmentation of more than one peptide ions to generate tandem spectra.[16]

The method to evaluate the statistical significance of the match between the observed and theoretical spectra could aid in augmenting the true positive rate of peptide detection. The k-permuted decoy database approach has been proven to increase the true positive rate of neuropeptide detection compared to approaches implemented in the database search programs.[88] The benefits of the Monte Carlo permuted decoy strategy to identify tryptic peptides have not been assessed. Unlike neuropeptides, tryptic peptides in general require trypsin digestion and show different fragmentation patterns due to presence of C-terminal basic residues. A public dataset of tryptic digest peptides from the plasma and serum proteins is available at (http://www.ludwig.edu.au).[20] The proteins were digested with trypsin enzyme and resulting peptides were analyzed using LCQ Deca XP ion trap mass spectrometer (Thermo-Finnigan, San Jose, CA, USA). In this dataset peptide assignments to 671 tandem mass spectra identified by the seven programs were independently validated by the experts in different laboratories.[20] The known peptide assignments were used in this study to evaluate the performance of the OMSSA and k-permuted decoy database.

The overall goal of this study is to evaluate the relative advantages of the k-permuted decoy database approach to identify tryptic peptides. Supporting aims were: (1) to evaluate the effectivity of the k-permuted decoy database approach to effectively use larger tryptic MS/MS datasets; (2) to

evaluate the ability of the k-permuted decoy database approach to discriminate between correct and incorrect peptide assignments on MS/MS datasets in which many spectra are either missed or incorrectly matched; (3) to evaluate the performance of the k-permuted decoy database approach and standard approach in OMSSA using commonly used the target-decoy search strategy to estimate False Discovery Rate (FDR); and (4) to evaluate the performance of the approaches using conventional Receiver Operating Characteristics (ROC) curves.

4.4 MATERIALS AND METHODS

DATASET AND TARGET DATABASE

The performance of the k-permuted decoy database approach was evaluated on annotated experimental tandem mass spectra obtained from the Human Plasma Proteome Project samples.[20] The plasma and serum proteins were digested with trypsin and resulting peptide mixture was separated using liquid chromatography (Agilent 1100 capillary column). The peptides were analyzed on an LCQ Deca XP ion trap mass spectrometer (Thermo-Finnigan, San Jose, CA, USA) using electrospray ionization source. This dataset with validated set of peptide assignments was downloaded from http://www.ludwig.edu.au and consisted of 5,806 tandem mass spectra (.dta format) that were analyzed using seven programs including Mascot, SEQUEST, PeptideProphet, Sonar, X!Tandem, Spectrum Mill, and Spectrum Mill (tag).[20] The dataset consisted of 671 annotated and 5,135 unannotated spectra. The known peptide identities for the 671 spectra ranged from 5 to 41 amino acids in length. Of the 671 spectra: 218, 360, and 93 had precursor charge

states of +1, +2, and +3, respectively. The dta spectral files were converted in to a single mascot generic format (mgf) file using a python script.

The spectra were searched against a target protein sequence database containing 68,711 protein entries downloaded from the RefSeq (ftp://ftp.ncbi.nih.gov/refseq; release 61 September 9, 2013). A reversed decoy database was also created by reversing the protein sequences in the target database. These reversed decoy sequences were appended at the end of target database for the combined target-decoy search strategy to estimate the false discovery rate.

DATABASE SEARCH PROGRAM AND DATABASE SEARCH STRATEGY

OMSSA (version 2.1.8), an open source program was used for the identification of peptides and to evaluate the performance of the k-permuted decoy database. This program was selected due to its relatively better performance for the small peptides and spectra containing noise.[16, 88] The program reported the number of matched ions between observed and theoretical spectra and significance values (*E-value* and *p-value*) for the PSMs. OMSSA uses a Poisson parametric distribution to compute significance values. The source code of OMSSA was modified to obtain the effective database size and Lambda (parameter for the Poisson distribution) for each spectrum. In this study, the OMSSA's *E-value* indicator of the match quality was evaluated. The *E-value* indicator was selected based on previous studies (chapters II and III) in which *E-value* outperformed the other indicators reported by OMSSA.

The following search specifications were used for the database searches: (1) precursor mass tolerance: 3.0 Da; (2) fragment mass tolerance: 0.5 Da; (3) variable PTMs: Oxidation of

Methionine and Carbamylation of Lysine; (4) mass type: monoisotopic; (5) enzyme: "trypsin" for the forward and reverse protein sequence databases or "whole" option to prevent cleavage of the provided k-permuted decoy database; (6) Digestion: partial digestion of protein sequence database that allows one non-tryptic termini in the resulting peptides; (7) maximum number of missed cleavages: 1; (8) minimum peptide length: 4; (9) minimum number of m/z peaks a spectrum must have: 2; (10) Option "ht": at-least one of theoretical spectra peak must match to one of the eight most intense experimental spectra peak; and (11) the upper limit on the maximum peptide length (i.e., 40 amino acids) for the semi-tryptic digestion was disabled.

GENERATION OF PERMUTED DECOY DATABASE

The Monte Carlo permutation approach was used to generate k-permuted decoy peptides by randomly sampling a subset (k) of decoy peptides for each spectrum. The decoy peptides are sequences of amino acids that are not present in the target database. The match of a spectrum against decoy peptides is considered incorrect and can be used to generate a reference null distribution to assign significance values to the peptide-spectrum matches. For peptide-spectrum matches, permutation *p-values* were computed using a k-permuted decoy database that indicated the probability that a match between peptide and spectra is due to chance. In a recent study, the performance of the permuted approach was evaluated using Whole sequence and Ends k-permuted decoy databases that provided better sensitivity and discrimination in the performance of the different scores or indicators reported by the three database search programs, respectively.[88]

In the Whole sequence k-permuted decoy databases, the k number of complete peptide sequences that were within 3.0 Da of the spectrum masses were randomly generated without replacement using a python script. For each position in the decoy peptides, the amino acids were randomly sampled with replacement from a list of 19 standard amino acids. The leucine and isoleucine were treated as isobaric amino acids and only Leucine was used to generate decoy peptides. For each experimental spectrum a separate decoy database was created containing 1×10^5 decoy peptides. The consideration of separate decoy databases for each spectrum can reduce the overall search time when multiple database search programs are considered. This strategy of permuted decoy database creation can be applied to any type of peptides, experiments, and database search programs.

The tandem mass spectra were searched against the k-permuted decoy peptide databases using OMSSA with enzymatic settings that prevented the cleavage of the provided peptide sequences. For each spectrum, the peptide matches in the k-permuted database that were identical in terms of match scores (i.e., Poisson Lambda, the number of matched ions, *p-value*, and *E-value*) and had masses within 3 Da of each other were treated as homeometric[86] matches. The homeometric matches were counted only once in calculation of permutation *p-values*. The problem of zero *p-values* was avoided by using the following formula for the *p-value* computation with plus one representing target peptide score that is always equal to itself:[73]

$$P(perm) = \frac{1 + \sum_{i=1}^{N} t(r) \geq t(s)}{N + 1}$$

Where *t(r)* is the score for the permuted peptides, *t(s)* is the score for the target peptides, and *N* is the effective database size for each particular spectrum. The two-fold search strategy was implied.

a) Tandem mass spectra were searched against the target database using OMSSA and the permutation *p-values* were computed for the OMSSA's *E-values* assigned to each peptide-spectrum match using k-permuted decoy databases.

b) Tandem mass spectra were searched against the concatenated target-reversed database using OMSSA and the permutation *p-values* were calculated for the OMSSA's *E-values* of the peptide-spectrum matches using k-permuted decoy databases.

PERFORMANCE EVALUATIONS: FDR AND ROC CURVES

For the two strategies, the performance (i.e., peptide detection at a specific FDR-based threshold) of the OMSSA *E-value* and permutation approach was compared at a 5% FDR. Only the best hit for each spectrum was considered in the FDR calculation and performance evaluation. The peptide-spectrum matches were arranged in the increasing order of their *E-values* and permutation *p-values* to compute the FDR. The significance value at which FDR was immediately below 5% was selected as thresholds.

For the target database search strategy, FDR was calculated with the assumption that the incorrect hits in the forward database are already known. This was accomplished using 671 annotated spectra with known sequence identities, while the remaining 5,135 spectra with

unknown peptide sequences were not considered in FDR calculation. The FDR was estimated with the following formula:

$$FDR = \frac{FP}{TP + FP} \qquad (1)$$

Where FP (false positive) and TP (true positives) were the number of incorrect and correct peptides hits with significance values below the threshold (i.e., significant matches), while TN (true negatives) and FN (false negatives) were the number of incorrect and correct peptide hits with significance values above the threshold (i.e., insignificant matches).

For the concatenated target-reversed database, FDR was calculated with the assumption that the incorrect hits in the forward database are not known (common case in MS/MS proteomics). All 5,806 spectra were considered in the FDR calculation using the following formula:

$$FDR = \frac{number\ of\ decoys}{number\ of\ targets} \qquad (2)$$

Where, number of decoys and targets represent the number of reversed and target peptides receiving significance values more significant than the threshold. The selected FDR formula produces less conservative FDR estimates than the *(2.decoys) / (target + decoy)* formula.[92] This procedure allowed the classification of 671 annotated spectra into TP, FN, TN, and FP at a particular score threshold.

Standard ROC curves were also used to compare the performance of the OMSSA's *E-value* and permutation approach. ROC curves were generated in OriginPro (version 8.6; http://www.originlab.com/). Plots showed tradeoff between Sensitivity (fraction of correct

significant peptide identifications among all correct identifications) and 1-specificity (fraction of incorrect significant identifications among all incorrect identifications) across different significance thresholds and provided a useful way to compare performance of the two approaches.

4.5 RESULTS AND DISCUSSION

The sensitivity (i.e., number of correct peptides matched at a given FDR threshold), and the significance levels of the peptide assignments to validated tandem mass spectra of the tryptic peptides obtained from the serum and plasma protein samples were evaluated. The peptide matches were filtered with 5% FDR, where FDR was calculated from the best hits for the tandem mass spectra matching in the target database and target-reversed concatenated databases. The peptide identification rates were compared between the OMSSA's *E-value* indicator (best indicator for the OMSSA based on previous two chapters) and permuted significance levels attained from the k-permuted decoy database using *E-value* indicator. The performance of the two approaches was compared in the target database and target-reversed concatenated database. The results were further verified by conventional ROC curves that highlighted the performance (i.e., tradeoff between Sensitivity *vs.* 1-Specficity) of the OMSSA's *E-value* and permutation approach regardless of any specific threshold based on false positive rate.

SENSITIVITY OF THE OMSSA'S *E-VALUE* IN THE STANDARD TARGET DATABASE

Sensitivity is a measure of the ability of the database search program to correctly match tandem mass spectra to peptide sequences in the database.[20] **Table 4.1** summarizes the number of annotated spectra that provided correct and incorrect matches in the target database.

The known peptide identities of the 671 annotated tandem spectra in this dataset were compared against the top hit assigned to each spectrum by OMSSA to classify them as correct and incorrect matches. In the target database, OMSSA correctly matched 469 (69.90%) peptides irrespective of the match significance levels. Of the remaining 202 incorrectly matched tandem mass spectra, 49 spectra were not detectable for the OMSSA either due to the absence of candidate peptides from the RefSeq database (39) or search parameter settings (10 spectra with N-terminal carbamylation). Any peptide match to these spectra was treated either as false positive or true negative identification based on significance threshold during the estimation of FDR. Spectra with no candidate peptides in the RefSeq database correspond to immunoglobulin genes that undergo extensive rearrangement and protein RefSeqs are not available for these genes. The remaining 153 spectra provided incorrect matches even in the presence of candidate peptides in the target database. Most of these spectra were incorrect either due to low intensity of the signal peaks relative to the noise peaks (high intensity noise peaks get preference over low intensity signal peaks due to filtering steps of OMSSA) or presence of better scoring modified peptides.

The sensitivity of the OMSSA's *E-value* was examined at a 5% FDR. In MS/MS based analysis, usually the incorrect matches in the target database are not known and the FDR is estimated using target-decoy approach or mixture model approach.[41] However, due the presence of annotated spectra in the dataset, the FDR is calculated directly as the proportion of incorrectly matched annotated spectra in the target database using formula one (see materials and methods section). For the calculation of FDR, the peptide matches were arranged in the ascending order of OMSSA's *E-value* and permutation *p-values*. The value at which the FDR was immediately below 5% was selected as the threshold.

Table 4.2 summarizes the number of true positive, false negative, true negative, false positive identifications at a 5% FDR for the OMSSA's *E-value* in the target database. OMSSA significantly detected 384 (81.88%) peptide identifications out of 469 at s 5% FDR. The 85 peptides that were not significantly detected had precursor charge state of +1 (54 peptides) and +2 (31), while all correctly matched peptides with +3 precursor charge state were significantly detected at 5% FDR. The false negative peptides had *E-values* > 1×10^{-1} and less number of matched fragment ions (<13) which made them indistinguishable from other incorrect matches in the target database. This was consistent with the previous study indicating that with insufficient number of matched fragment ions (either due to missing ions, small peptide size etc.,) the significance values of the OMSSA tends to increase (i.e., become less significant).[16] Of the 85 insignificant peptides, 100%, 100%, 94%, 62%, 45%, 41%, 20%, and 25% had 5, 6, 7, 8, 9, 10, 11, and 12 number of matched fragment ions. False negative peptides showed more overlap with the true negatives and false positives in terms of OMSSA *E-values* which made those peptides insignificant using the OMSSA *E-values*.

SENSITIVITY OF THE PERMUTATION APPROACH IN STANDARD TARGET DATABASE

The *E-value* indicator was considered in this study due to its higher sensitivity and discriminatory power than the other peptide-spectrum match quality indicators in the previous studies (chapter II and III). OMSSA's *E-values* of around 37% correctly matched annotated peptides were more significant than the ones obtainable with the k-permuted decoy database size used in this study. These 37% peptides were also significant with the k-permuted decoy database.

Overall, k-permuted showed better performance for the false negative identifications produces by the standard OMSSA's *E-value* due to overlap with true negatives and false positives.

Table 4.2 summarizes the number of TP, FN, TN, FP identifications at a 5% FDR for the k-permuted decoy database. At a 5 % FP rate, the k-permuted decoy database identified 417 (88.91%) of the annotated peptides correctly matched by the OMSSA database search program. The sensitivity of the k-permuted decoy database improved with the increasing number of matched fragment ions. Peptide detection rate was 0%, 42%, 12%, 52%, 68%, 79%, 96%, 97%, and 100% for peptides with 5, 6, 7, 8, 9, 10, 11, 12, and 13 matched fragment ions, respectively. All peptides with +3 precursor charge state were significantly detected, while 38 and 14 peptides were missed by the k-permuted decoy database with precursor charge states of +1 and +2, respectively.

Consistent with our previous studies, accurate permutation *p-values* for the *E-value* indicator could be estimated with smaller number of permutations ($\approx 10^5$). The large *E-values* (less significant) assigned to the peptide matches in the target database allowed more decoy peptides to receive equal or more extreme values than the target peptides even with less number of permutations, thus separating correct and incorrect matches in better fashion. This is because the target peptide matches with less number of matched ions and less intense fragment ion peaks can receive higher *E-values* which in turn can make target database matches with extremely large *E-values* non-significant.

COMPARISON OF PERFORMANCE IN THE TARGET DATABASE

Table 4.3 summarizes the $\log_{10}$-transformed significance levels of the OMSSA's *E-values* for the peptides that were significantly detected by the OMSSA *E-value* and k-permuted decoy database, k-permuted decoy database only, OMSSA *E-value* only, and not detected by any approach.

Comparison of the peptide identifications between the OMSSA *E-value* and k-permuted decoy database indicated that 378, 39, and 6 peptides were significantly detected by the two approaches, only k-permuted decoy database, and only OMSSA *E-value*, respectively. Consensus among approaches was lower for the small and large peptide matches with fewer than 13 matched fragment ions. The OMSSA *E-value* and k-permuted decoy database detected 0%, 0%, 0.06%, 31%, 45%, 55%, 80%,75%, 93% peptides with 5, 6, 7, 8, 9, 10, 11, 12, and 13 number of matched fragment ions. The 46 peptides were not detected by any approach. The peptides that were not significant across both approaches had less number of matched ions (ranging between 5 to 12) and large OMSSA's *E-values* (mean=23.89, std. dev=51). Due to large *E-values* and less number of matched ions the more permuted decoy peptides had extreme *E-values* than the correct matches. The peptides with extreme *E-values* (especially short peptides with +1 charge state) were not distinguishable from other incorrect matches in the target database.

Figure 4.1 depicts the ROC curve comparison between the OMSSA's *E-value* indicator and permutation *p-value* approach. Roc curves shows the effectiveness of each match indicators across different significance thresholds by allowing investigation of tradeoff between true positive rate (sensitivity) and false positive rate (1-specificity). The diagonal line from lower left corner to

the upper left corner in the curve indicates the usefulness of the two approaches. The curves above the diagonal lines denote that the discrimination between correct and incorrect peptide identifications provided by the approaches was not due to chance. The area under the curve (AUC) given in plot provides information about the discriminatory power (value close to one shows good power) of each approach across different significance thresholds.

The curves depicts the tradeoff between the true positive rate and false positive rate across different thresholds for the OMSSA's *E-value* indicator and permutation *p-value* calculated from the *E-value* indicator using k-permuted decoy database. Overall, the k-permuted decoy database performed better than the OMSSA *E-value* in the target database. The AUC indicating effectiveness for the k-permuted decoy database and OMSSA *E-value* in separating correct from incorrect matches was around 0.95 and 0.94, respectively. The false positive rate of the OMSSA *E-value* was slightly lower than the k-permuted decoy database when the sensitivity was below 68%. However, false positive rate of the OMSSA *E-value* increased with slight increase in sensitivity of the peptide identifications. The performance of k-permuted decoy database and OMSSA *E-value* converged around sensitivity of 79%. The k-permuted decoy database achieved higher sensitivity (around 89%) while keeping false positive rate below 5% than OMSSA *E-value* with sensitivity around 82%.

SENSITIVITY OF THE PEPTIDE IDENTIFICATION USING TARGET-REVERSE APPROACH

In MS/MS analysis, target-decoy search strategy is commonly used to estimate the FDR as the peptide identities for the incorrect matches in the target database are not known. The target-

decoy strategy is easy to implement and is based on the assumption that the incorrect matches in the target database and decoy database have identical score distributions. In this study, protein sequences in the target database were reversed and appended to the target database to conduct combined target-decoy database search. This is because the combined search strategy is considered to produce less conservative FDR estimates.[92] The 5,806 spectra were searched against the combined database and using the best hit for each spectrum was considered. The permutation *p-values* for the OMSSA *E-values* were calculated using k-permuted decoy database as the ratio of permuted peptides receiving *E-values* as extreme as the *E-value* of the hits in the target-decoy database. The performance of the OMSSA *E-value* and k-permuted decoy database was compared in terms of sensitivity and overlap among peptide identifications in the target and permuted databases at 5% FDR, where FDR was calculated as the ratio of decoy and target matches at different significance threshold.

Table 4.1 summarizes the number of annotated spectra with correct and incorrect peptide assignments in the target-reverse decoy database across three precursor charge states. OMSSA correctly matched 459 (68.41%) annotated spectra. The target-reverse database reduced the number of correct peptide identifications by 1% relative to the target only database. Seven peptides with +1 charge state that were not correctly matched by the OMSSA in the target-reverse database were not significantly detected by the OMSSA in the target database at 5% FDR. This could be one of the reasons for the apparent reduction in the number of false negative peptides in the target-reverse database relative to the target only database in addition to the increase in number of available spectra for the calculation of FDR.

Table 4.4 summarizes the number of TP, FN, TN, and FP matches for the 671 annotated spectra using OMSSA's *E-value* and k-permuted database. The OMSSA *E-value* significantly detected 84.10% peptides out of the 459 correctly annotated spectra at 5% FDR based threshold. Consistent with the target database results, about 92% true positives matches had *E-values* < 1 x 10^{-1}, while false negatives had *E-values* > 1 x 10^{-1} and showed more overlap with the true negatives and false positives. The k-permuted database detected 87.15% of the correctly matched annotated peptides. Compared with the OMSSA's *E-value*, the k-permuted database had higher sensitivity and specificity (33% less false positives). Consistent with the target database, k-permuted approach detected more peptides that showed higher overlap with false positives and true negatives using OMSSA's *E-values*. Furthermore, OMSSA's *E-value* and k-permuted decoy database also significantly detected 566 and 594 unannotated spectra which could be either false positives or true positives but there is no assurance about these results.

4.6 CONCLUSIONS

The present study demonstrated that the k-permuted decoy database can be used for both peptide matches coming from the target database and commonly used target-decoy database search strategy. The peptide-spectrum matches were filtered with false discovery rate calculated in two different ways. The results indicated that in both databases the k-permuted database had higher specificity (less false positives) than the standard approach implemented in the OMSSA program.

The k-permuted decoy database allowed detection of more annotated peptides in both the target database and target-reverse database relative to the OMSSA *E-value* at a fixed false discovery rate of 5%. Better detection rate was achieved by separating borderline correct matches

with large *E-values* (less significant) from incorrect matches. The area under the ROC curves was slightly better for the k-permuted decoy database. The study demonstrated that the k-permuted decoy database can be integrated with any database search program and current standards in MS/MS based analysis.

4.7 FIGURES

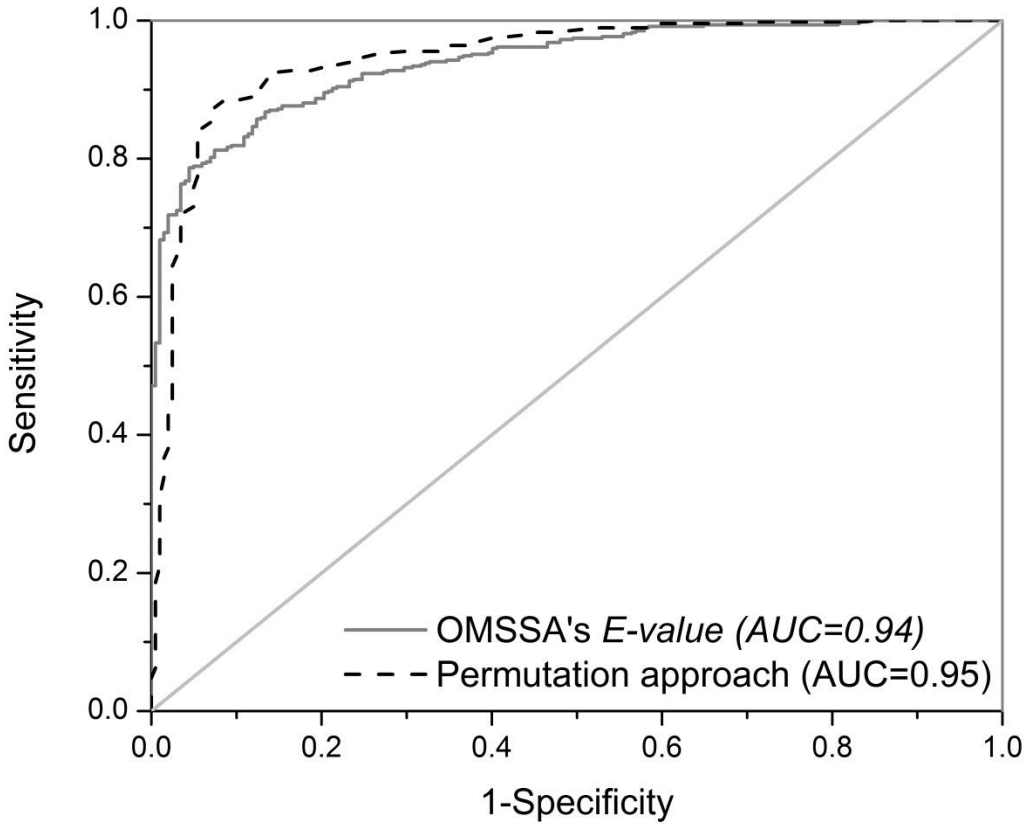

Figure 4.1. ROC curves for match indicators in the target and permuted databases. Plot compares discriminatory powers of target database versus permuted database using permuted *p-values* from the OMSSA's *E-value* indicators against target database *E-value*.

4.8 TABLES

Table 4.1. The number of correctly and incorrectly matched annotated spectra in the target and concatenated target-reverse databases, irrespective of the match significance levels across three precursor charge states.

Peptide hits for the annotated spectra	Target database				Concatenated target-reverse database			
	Total spectra	Precursor charge state distribution			Total spectra	Precursor charge state distribution		
		+1	+2	+3		+1	+2	+3
Matched	469	98	303	68	459	92	300	67
Misidentified	202	120	57	25	212	126	60	26
Total	671	218	360	93	671	218	360	93

Table 4.2. Sensitivity of the OMSSA's *E-value* and k-permuted decoy database at 5% False Discovery Rate when spectra were searched against standard target database.

Approach[a]	Total spectra	5% False Discovery Rate (FDR)[b]				
		TP[c]	FN	TN	FP	Threshold
OMSSA's *E-value*	671	384	85	182	20	0.21557
Permutation	671	417	52	181	21	1.995e-5

[a] *E-value*: OMSSA's *E-value*; Permutation: *p-value* computed using OMSSA's *E-value*;
[b] FDR: FP/TP+FP;
[c] TP: number of correctly matched peptides significant at 5% FDR; FN: number of correctly matched peptides not significant at 5% FDR; TN: number of incorrectly matched peptides not significant at 5% FDR; FP: number of incorrectly matched peptides significant at 5% FDR; Threshold: significance threshold for the 5% FDR.

Table 4.3. The number of spectra significantly detected by both OMSSA's *E-value* and k-permuted decoy database, k-permuted decoy database only, OMSSA's *E-value* only, and not detected by any approach in the target database at a false discovery rate of 5%.

Log_{10} *E-value*	**Permuted & OMSSA's *E-value***	**Permuted approach only**	**OMSSA's *E-value* only**	**Not significant in both approaches**
-3	0	0	0	0
-2	0	0	0	3
-1	0	3	0	13
0	13	36	1	30
1	34	0	3	0
2	46	0	2	0
3	39	0	0	0
4	36	0	0	0
5	36	0	0	0
6	174	0	0	0

Table 4.4. Sensitivity of the OMSSA's *E-value* and k-permuted decoy database at 5% False Discovery Rate when spectra were searched against target-reverse combined database.

Approach[a]	Total spectra	Target database matches[b]	Reverse database matches[c]	5% False Discovery Rate (FDR)[d]				
				TP[e]	FN	TN	FP	Unknown
OMSSA's *E-value*	5806	3834	1972	386	73	173	39	566
Permutation	5806	3834	1972	400	59	186	26	594

[a] *E-value*: OMSSA's *E-value*; Permutation: *p-value* computed using OMSSA's *E-value*;
[b] The number of spectra with peptide matches from the target database.
[c] The number of spectra with peptide matches from the reverse database.
[d] FDR: #reverse/#targets; based on the assumption that usually incorrect matches in the target database are not known, the unknown spectra with match in target database were treated as target while calculating FDR.
[e] TP: number of correctly matched peptides significant at 5% FDR; FN: number of correctly matched peptides not significant at 5% FDR; TN: number of incorrectly matched peptides not significant at 5% FDR; FP: number of incorrectly matched peptides significant at 5% FDR; Unknown: unannotated spectra those were significant at 5% FDR; Threshold: significance threshold for the 5% FDR.

CHAPTER V: CONCLUSION

The database search programs are commonly used to identify neuropeptides and proteins in tandem mass spectrometry experiments. In MS/MS analysis, typical challenges for the database search programs include failure to correctly identify tandem mass spectra and discriminate correct from incorrect identifications. Previous studies have shown that the programs designed for protein identification (usually one unique peptide matched with high confidence is enough to identify protein) are not optimized for the identification of neuropeptides and short peptides present in biological sample.[16] For the ideal simulated tandem spectra containing all possible fragment ions, the database search programs can correctly assign peptide sequences to most tandem mass spectra. However, accurate assignment of the match significance levels remains challenging due to spectral quality issues and limitations of the parametric distribution approaches in programs for the shorter peptides.

In my studies, permutation testing was used to overcome the limitations associated with the parametric approaches already implemented in the database search programs. Three studies were conducted with the overall aims: (a) to develop and integrate permutation resampling approach with the database search programs; (b) to identify the match indicators that provide optimal performance within and across programs; and (c) to evaluate the classification performance of the approach relative to the already implemented approaches in the database search programs. The permutation databases based on the complete permutations and terminal permutations of the peptide sequences were developed and performance of various scores within and across programs

was tested. The results provided us an insight about suitability of the various scores in each program to be used for computation of significance values.

Resampling approach can be used to validate the suitability of the significance estimation approached implemented in the database search programs. The magnitude of the scores is reduced by factors such as spectral complexity and low signal-to-noise ratio even in presence of correct matches, which in turn have an effect on the match significance values. These spectral quality issues had larger impact on the significance estimation approaches already implemented in the programs relative to the resampling approach. The approach improved peptide detection rate in Crux and X! Tandem programs which underestimated significance values with parametric approaches leading to reduction in the peptide identifications. Furthermore, the permutation testing validated the suitability of the significance estimation approach implemented in OMSSA for the neuropeptides.

Additionally, the resampling approach provides comparable basis for benchmarking various database search programs. The same peptide identifications from the different database search programs are less likely to be comparable due to difference in significance estimation approaches using standard ways to adjust for the multiple hypotheses testing. A common approach in MS/MS studies is to use target-decoy database search strategy to compute significance thresholds based on certain level of false discovery rate to control for multiple hypothesis testing. On the other hand, resampling approach allows computation of significance levels for various programs and scores in the same fashion, which allows their benchmarking using standard threshold adjustment procedures for multiple hypothesis testing. This can be useful to save

computational search time which increases when spectra are searched against target-decoy databases.

Consensus approach (use of multiple programs) assures that the peptide matches are less likely to be false positives. Our studies demonstrate that the programs show less agreement for the peptide identifications at more stringent thresholds with already implemented significance estimation approaches. The resampling approach improves consensus among programs due to higher peptide detection rate for the Crux and X! Tandem that previously underestimated significance values. Furthermore, the resampling approach confirms the previous notion that some scoring functions are better suited for peptide identification than other programs.

The database search programs assign peptide matches to tandem mass spectra even when appropriate candidate peptides are missing from the database. This requires that significance estimation approaches should be able to provide good separation between correct and incorrect peptide identifications. The results demonstrated that the resampling approach performs better in discriminating correct peptides matches with less than 13 matched fragment ions from other incorrect matches.

The permutation testing provides more accurate estimates of the null distribution, where the null distribution is formed by either complete random peptides or terminal permutations of the peptides that fall within a certain mass tolerance range of each spectrum. The two types of permuted decoy databases provided enough number of target alike decoy peptides to discriminate the performance of poor match indicators from the strong indicators of the peptide match. Some scores or indicators within and across programs had lower peptide detection rate regardless of the

type of the permuted decoy database. Terminal permutations generated more decoy peptides that can receive more extreme scores than the target database matches which in turn reduced the peptide detection using indicators within and across database search programs.

The availability of the open source database search programs to proteomics and bioinformatics community is important to evaluate and refine their scoring functions. The ability to retrieve and test a set of intermediate and final match indicators from the database search programs provides useful information about the relative strengths of various match indicators within and across programs. Model free property of the k-permuted decoy database allows consideration of the multiple match indicators in confidently identifying peptides that are borderline significant. Furthermore, the sharing of source code to generate k-permuted decoy database will be beneficial for the proteomics community to further explore the biological, statistical, and computational basis of the approach for the other database search programs and spectral identification approaches.

Future studies are needed to uncover the impact of target peptide amino acid composition on the significance levels estimation using k-permuted decoy databases. Furthermore, a study about assigning significance values to protein matches rather than the spectrum matches needs to be undertaken.

CHAPTER VI: REFERENCES

1. Hook, V.; Funkelstein, L.; Lu, D.; Bark, S.; Wegrzyn, J.; Hwang, S. R. Proteases for processing proneuropeptides into peptide neurotransmitters and hormones, *Annu. Rev. Pharmacol. Toxicol.* **2008**, 393-423.

2. Bora, A.; Annangudi, S. P.; Millet, L. J.; Rubakhin, S. S.; Forbes, A. J.; Kelleher, N. L.; Gillette, M. U.; Sweedler, J. V. Neuropeptidomics of the supraoptic rat nucleus, *J. Proteome Res.* **2008**, *11*, 4992-5003.

3. Svensson, M.; Skold, K.; Nilsson, A.; Falth, M.; Nydahl, K.; Svenningsson, P.; Andren, P. E. Neuropeptidomics: MS applied to the discovery of novel peptides from the brain, *Anal. Chem.* **2007**, *1*, 15-6, 18-21.

4. Strand, F. L. *Neuropeptides: regulators of physiological processes;* Cambridge, Mass. : MIT Press, ©1999: 1999.

5. von Eggelkraut-Gottanka, R.; Beck-Sickinger, A. G. Biosynthesis of peptide hormones derived from precursor sequences, *Curr. Med. Chem.* **2004**, *20*, 2651-2665.

6. Svensson, M.; Skold, K.; Svenningsson, P.; Andren, P. E. Peptidomics-based discovery of novel neuropeptides, *J. Proteome Res.* **2003**, *2*, 213-219.

7. Tegge, A. N.; Southey, B. R.; Sweedler, J. V.; Rodriguez-Zas, S. L. Comparative analysis of neuropeptide cleavage sites in human, mouse, rat, and cattle, *Mamm. Genome* **2008**, *2*, 106-120.

8. Naggert, J. K.; Fricker, L. D.; Varlamov, O.; Nishina, P. M.; Rouille, Y.; Steiner, D. F.; Carroll, R. J.; Paigen, B. J.; Leiter, E. H. Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity, *Nat. Genet.* **1995**, *2,* 135-142.

9. Lloyd, D. J.; Bohan, S.; Gekakis, N. Obesity, hyperphagia and increased metabolic efficiency in Pc1 mutant mice, *Hum. Mol. Genet.* **2006**, *11,* 1884-1893.

10. Fricker, L. D. Neuropeptide-processing enzymes: applications for drug discovery, *AAPS J.* **2005**, *2,* E449-55.

11. Bendena, W. G.; Campbell, J.; Zara, L.; Tobe, S. S.; Chin-Sang, I. D. Select Neuropeptides and their G-Protein Coupled Receptors in Caenorhabditis Elegans and Drosophila Melanogaster, *Front. Endocrinol. (Lausanne)* **2012**, 93.

12. Boonen, K.; Landuyt, B.; Baggerman, G.; Husson, S. J.; Huybrechts, J.; Schoofs, L. Peptidomics: the integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis, *J. Sep. Sci.* **2008**, *3,* 427-445.

13. Aebersold, R.; Mann, M. Mass spectrometry-based proteomics, *Nature* **2003**, *6928,* 198-207.

14. Yates, J. R.; Ruse, C. I.; Nakorchevsky, A. Proteomics by mass spectrometry: approaches, advances, and applications, *Annu. Rev. Biomed. Eng.* **2009**, 49-79.

15. Chait, B. T. Chemistry. Mass spectrometry: bottom-up or top-down?, *Science* **2006**, *5796*, 65-66.

16. Akhtar, M. N.; Southey, B. R.; Andren, P. E.; Sweedler, J. V.; Rodriguez-Zas, S. L. Evaluation of database search programs for accurate detection of neuropeptides in tandem mass spectrometry experiments, *J. Proteome Res.* **2012**, *12*, 6044-6055.

17. Reid, G. E.; McLuckey, S. A. 'Top down' protein characterization via tandem mass spectrometry, *J. Mass Spectrom.* **2002**, *7*, 663-675.

18. LeDuc, R. D.; Taylor, G. K.; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli, J. S.; Kelleher, N. L. ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry, *Nucleic Acids Res.* **2004**, *Web Server issue*, W340-5.

19. Resing, K. A.; Ahn, N. G. Proteomics strategies for protein identification, *FEBS Lett.* **2005**, *4*, 885-889.

20. Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S.; Simpson, R. J. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis, *Proteomics* **2005**, *13*, 3475-3490.

21. Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry, *J. Am. Chem. Soc.* **1999**, *4*, 806-812.

22. Nesvizhskii, A. I. Protein identification by tandem mass spectrometry and sequence database searching, *Methods Mol. Biol.* **2007**, 87-119.

23. UniProt Consortium The Universal Protein Resource (UniProt), *Nucleic Acids Res.* **2007**, *Database issue,* D193-7.

24. UniProt Consortium The Universal Protein Resource (UniProt) in 2010, *Nucleic Acids Res.* **2010**, *Database issue*, D142-8.

25. Falth, M.; Skold, K.; Norrman, M.; Svensson, M.; Fenyo, D.; Andren, P. E. SwePep, a database designed for endogenous peptides and mass spectrometry, *Mol. Cell. Proteomics* **2006**, *6*, 998-1005.

26. Falth, M.; Svensson, M.; Nilsson, A.; Skold, K.; Fenyo, D.; Andren, P. E. Validation of endogenous peptide identifications using a database of tandem mass spectra, *J. Proteome Res.* **2008**, *7*, 3049-3053.

27. Stein, S. E.; Rudnick, P. A. NIST Peptide Tandem Mass Spectral Libraries. Human Peptide Mass Spectral Reference Data, *H. sapiens*, ion trap, Official Build Date: Feb. 4, 2009. National Institute of Standards and Technology, Gaithersburg, MD, 20899. *http://peptide.nist.gov* (accessed February 28, 2012).

28. Kim, Y.; Bark, S.; Hook, V.; Bandeira, N. NeuroPedia: neuropeptide database and spectral library, *Bioinformatics* **2011**, *19*, 2772-2773.

29. Wang, J.; Perez-Santiago, J.; Katz, J. E.; Mallick, P.; Bandeira, N. Peptide identification from mixture tandem mass spectra, *Mol. Cell. Proteomics* **2010**, *7*, 1476-1485.

30. Liu, F.; Baggerman, G.; Schoofs, L.; Wets, G. The construction of a bioactive peptide database in Metazoa, *J. Proteome Res.* **2008**, *9*, 4119-4131.

31. Liu, F.; Baggerman, G.; Schoofs, L.; Wets, G. Uncovering conserved patterns in bioactive peptides in Metazoa, *Peptides* **2006**, *12*, 3137-3153.

32. B. R. Southey, M. N. Akhtar, P. E. Andrén, J. V. Sweedler and S. L. and Rodriguez-Zas. A comprehensive resource in support of sequence-based studies of neuropeptides, **2013**, 144.

33. Burbach, J. P. Neuropeptides from concept to online database www.neuropeptides.nl, *Eur. J. Pharmacol.* **2010**, *1*, 27-48.

34. Zamyatnin, A. A.; Borchikov, A. S.; Vladimirov, M. G.; Voronina, O. L. The EROP-Moscow oligopeptide database, *Nucleic Acids Res.* **2006**, *Database issue*, D261-6.

35. Pruitt, K. D.; Tatusova, T.; Brown, G. R.; Maglott, D. R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy, *Nucleic Acids Res.* **2012**, *Database issue*, D130-5.

36. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry, *J. Comput. Biol.* **1999**, *3-4*, 327-342.

37. Steen, H.; Mann, M. The ABC's (and XYZ's) of peptide sequencing, *Nat. Rev. Mol. Cell Biol.* **2004**, *9*, 699-711.

38. Chen, T.; Kao, M. Y.; Tepel, M.; Rush, J.; Church, G. M. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry, *J. Comput. Biol.* **2001**, *3*, 325-337.

39. Kapp, E.; Schutz, F. Overview of tandem mass spectrometry (MS/MS) database search algorithms, *Curr. Protoc. Protein Sci.* **2007**, Unit#25.2.

40. Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry, *Nat. Methods* **2007**, *10*, 787-797.

41. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics, *J. Proteomics* **2010**, *11*, 2092-2123.

42. Deutsch, E. W. Tandem mass spectrometry spectral libraries and library searching, *Methods Mol. Biol.* **2011**, 225-232.

43. Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm, *J. Proteome Res.* **2004**, *5*, 958-964.

44. Park, C. Y.; Klammer, A. A.; Kall, L.; MacCoss, M. J.; Noble, W. S. Rapid and accurate peptide identification from tandem mass spectra, *J. Proteome Res.* **2008**, *7*, 3022-3027.

45. Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra, *Bioinformatics* **2004**, *9*, 1466-1467.

46. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis* **1999**, *18*, 3551-3567.

47. Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, **1994**, 976-989.

48. Diament, B. J.; Noble, W. S. Faster SEQUEST searching for peptide identification from tandem mass spectra, *J. Proteome Res.* **2011**, *9*, 3871-3879.

49. Eng, J. K.; Searle, B. C.; Clauser, K. R.; Tabb, D. L. A face in the crowd: recognizing peptides through database search, *Mol. Cell. Proteomics* **2011**, *11*, R111.009522.

50. Cooper, B. The problem with peptide presumption and low Mascot scoring, *J. Proteome Res.* **2011**, *3*, 1432-1435.

51. Mann, M.; Jensen, O. N. Proteomic analysis of post-translational modifications, *Nat. Biotechnol.* **2003**, *3*, 255-261.

52. Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry, *Genome Res.* **2001**, *2*, 290-299.

53. Johnson, R. S.; Davis, M. T.; Taylor, J. A.; Patterson, S. D. Informatics for protein identification by mass spectrometry, *Methods* **2005**, *3*, 223-236.

54. Papayannopoulos, I. A. The interpretation of collision-induced dissociation tandem mass spectra of peptides, *Mass Spectrometry Reviews* **1995**, 49-73.

55. Marcotte, E. M. How do shotgun proteomics algorithms identify proteins?, *Nat. Biotechnol.* **2007**, *7*, 755-757.

56. Klammer, A. A.; Park, C. Y.; Noble, W. S. Statistical calibration of the SEQUEST XCorr function, *J. Proteome Res.* **2009**, *4*, 2106-2113.

57. Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes, *Anal. Chem.* **2003**, *4*, 768-774.

58. Frese, C. K.; Boender, A. J.; Mohammed, S.; Heck, A. J.; Adan, R. A.; Altelaar, A. F. Profiling of diet-induced neuropeptide changes in rat brain by quantitative mass spectrometry, *Anal. Chem.* **2013**, *9*, 4594-4604.

59. Creasy, D. M.; Cottrell, J. S. Error tolerant searching of uninterpreted tandem mass spectrometry data, *Proteomics* **2002**, *10*, 1426-1434.

60. Yu, L.; Tan, Y.; Tsai, Y.; Goodlett, D. R.; Polfer, N. C. On the relevance of peptide sequence permutations in shotgun proteomics studies, *J. Proteome Res.* **2011**, *5*, 2409-2416.

61. Diz, A. P.; Carvajal-Rodriguez, A.; Skibinski, D. O. Multiple hypothesis testing in proteomics: a strategy for experimental work, *Mol. Cell. Proteomics* **2011**, *3*, M110.004374.

62. Franceschi, P.; Giordan, M.; Wehrens, R. Multiple comparisons in mass-spectrometry-based -omics technologies, *TRAC-TRENDS IN ANALYTICAL CHEMISTRY* **2013**, *50*, 11-21.

63. Holm, S. A simple sequentially rejective multiple test procedure, *Scandinavian journal of statistics* **1979**, 65-70.

64. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society.Series B (Methodological)* **1995**, 289-300.

65. Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, *Nat. Methods* **2007**, *3*, 207-214.

66. Balgley, B. M.; Laudeman, T.; Yang, L.; Song, T.; Lee, C. S. Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy, *Mol. Cell. Proteomics* **2007**, *9*, 1599-1608.

67. Yadav, A. K.; Kumar, D.; Dash, D. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry, *J. Proteome Res.* **2011**, *5*, 2154-2160.

68. Elias, J. E.; Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics, *Methods Mol. Biol.* **2010**, 55-71.

69. Higdon, R.; Hogan, J. M.; Van Belle, G.; Kolker, E. Randomized sequence databases for tandem mass spectrometry peptide and protein identification, *OMICS* **2005**, *4*, 364-379.

70. Kall, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases, *J. Proteome Res.* **2008**, *1*, 29-34.

71. Phipson, B.; Smyth, G. K. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn, *Stat. Appl. Genet. Mol. Biol.* **2010**, Article39-6115.1585. Epub 2010 Oct 31.

72. Lai, Y. Conservative adjustment of permutation p-values when the number of permutations is limited, *Int. J. Bioinform Res. Appl.* **2007**, *4*, 536-546.

73. Knijnenburg, T. A.; Wessels, L. F.; Reinders, M. J.; Shmulevich, I. Fewer permutations, more accurate P-values, *Bioinformatics* **2009**, *12*, i161-8.

74. Hummon, A. B.; Amare, A.; Sweedler, J. V. Discovering new invertebrate neuropeptides using mass spectrometry, *Mass Spectrom. Rev.* **2006**, *1*, 77-98.

75. Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. ProSight PTM 2.0: improved protein identification and characterization for top down mass spectrometry, *Nucleic Acids Res.* **2007**, *Web Server issue,* W701-6.

76. Xie, F.; London, S. E.; Southey, B. R.; Annangudi, S. P.; Amare, A.; Rodriguez-Zas, S. L.; Clayton, D. F.; Sweedler, J. V. The zebra finch neuropeptidome: prediction, detection and expression, *BMC Biol.* **2010**, , 28-7007-8-28.

77. Zhang, X.; Petruzziello, F.; Zani, F.; Fouillen, L.; Andren, P. E.; Solinas, G.; Rainer, G. High identification rates of endogenous neuropeptides from mouse brain, *J. Proteome Res.* **2012**, *5*, 2819-2827.

78. Jia, C.; Lietz, C. B.; Ye, H.; Hui, L.; Yu, Q.; Yoo, S.; Li, L. A multi-scale strategy for discovery of novel endogenous neuropeptides in the crustacean nervous system, *J. Proteomics* **2013**, 1-12.

79. Southey, B. R.; Lee, J. E.; Zamdborg, L.; Atkins, N.,Jr; Mitchell, J. W.; Li, M.; Gillette, M. U.; Kelleher, N. L.; Sweedler, J. V. Comparing label-free quantitative peptidomics approaches to characterize diurnal variation of peptides in the rat suprachiasmatic nucleus, *Anal. Chem.* **2014**, *1*, 443-452.

80. Sadygov, R. G.; Yates, J. R.,3rd A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases, *Anal. Chem.* **2003**, *15*, 3792-3798.

81. Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A.; Working Group on Publication Guidelines for Peptide and Protein Identification Data The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data, *Mol. Cell. Proteomics* **2004**, *6*, 531-533.

82. Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J.; Pevzner, P. A. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search, *Mol. Cell. Proteomics* **2010**, *12*, 2840-2852.

83. Alves, G.; Ogurtsov, A. Y.; Yu, Y. K. RAId_aPS: MS/MS analysis with multiple scoring functions and spectrum-specific statistics, *PLoS One* **2010**, *11*, e15438.

84. Ernst, M. D. Permutation Methods: A Basis for Exact Inference, *Statistical Science* **2004**, , 676-685.

85. Southey, B. R.; Amare, A.; Zimmerman, T. A.; Rodriguez-Zas, S. L.; Sweedler, J. V. NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides, *Nucleic Acids Res.* **2006**, *Web Server issue*, W267-72.

86. Frank, A. M.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A.; Pevzner, P. A. De novo peptide sequencing and identification with precision mass spectrometry, *J. Proteome Res.* **2007**, *1*, 114-123.

87. Yin, P.; Bousquet-Moore, D.; Annangudi, S. P.; Southey, B. R.; Mains, R. E.; Eipper, B. A.; Sweedler, J. V. Probing the production of amidated peptides following genetic and dietary copper manipulations, *PLoS One* **2011**, *12*, e28679.

88. Akhtar, M. N.; Southey, B. R.; Andren, P. E.; Sweedler, J. V.; Rodriguez-Zas, S. L. Evaluation of significance level assignment of database search programs using Monte Carlo permutation approach*;* March 24-26, 2014.

89. Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *J. Am. Soc. Mass Spectrom.* **1994**, *11*, 976-989.

90. Salmi, J.; Nyman, T. A.; Nevalainen, O. S.; Aittokallio, T. Filtering strategies for improving protein identification in high-throughput MS/MS studies, *Proteomics* **2009**, *4*, 848-860.

91. Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, *Nat. Biotechnol.* **2008**, *12*, 1367-1372.

92. Jeong, K.; Kim, S.; Bandeira, N. False discovery rates in spectral identification, *BMC Bioinformatics* **2012**, , S2-2105-13-S16-S2. Epub 2012 Nov 5.